INFERENCE OF NONPARAMETRIC HYPOTHESIS TESTING ON HIGH DIMENSIONAL LONGITUDINAL DATA AND ITS APPLICATION IN DNA COPY NUMBER VARIATION AND MICROARRAY DATA ANALYSIS

by

KE ZHANG

B.S., Wuhan Univeristy, China, 1996
M.S., Kansas State University, 2004

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2008

Abstract

High throughput screening technologies have generated a huge amount of biological data in the last ten years. With the easy availability of array technology, researchers started to investigate biological mechanisms using experiments with more sophisticated designs that pose novel challenges to statistical analysis. We provide theory for robust statistical tests in three flexible models. In the first model, we consider the hypothesis testing problems when there are a large number of variables observed repeatedly over time. A potential application is in tumor genomics where an array comparative genome hybridization (aCGH) study will be used to detect progressive DNA copy number changes in tumor development. In the second model, we consider hypothesis testing theory in a longitudinal microarray study when there are multiple treatments or experimental conditions. The tests developed can be used to detect treatment effects for a large group of genes and discover genes that respond to treatment over time. In the third model, we address a hypothesis testing problem that could arise when array data from different sources are to be integrated. We perform statistical tests by assuming a nested design. In all models, robust test statistics were constructed based on moment methods allowing unbalanced design and arbitrary heteroscedasticity. The limiting distributions were derived under the nonclassical setting when the number of probes is large. The test statistics are not targeted at a single probe. Instead, we are interested in testing for a selected set of probes simultaneously. Simulation studies were carried out to compare the proposed methods with some traditional tests using linear mixed-effects models and generalized estimating equations. Interesting results obtained with the proposed theory in two cancer genomic studies suggest that the new methods are promising for a wide range of biological applications with longitudinal arrays.

INFERENCE OF NONPARAMETRIC HYPOTHESIS TESTING ON HIGH DIMENSIONAL LONGITUDINAL DATA AND ITS APPLICATION IN DNA COPY NUMBER VARIATION AND MICROARRAY DATA ANALYSIS

by

KE ZHANG

B.S., Wuhan University, China, 1996
M.S., Kansas State University, 2004

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2008

Approved by:

Major Professor
Haiyan Wang

Copyright

Abstract

High throughput screening technologies have generated a huge amount of biological data in the last ten years. With the easy availability of array technology, researchers started to investigate biological mechanisms using experiments with more sophisticated designs that pose novel challenges to statistical analysis. We provide theory for robust statistical tests in three flexible models. In the first model, we consider the hypothesis testing problems when there are a large number of variables observed repeatedly over time. A potential application is in tumor genomics where an array comparative genome hybridization (aCGH) study will be used to detect progressive DNA copy number changes in tumor development. In the second model, we consider hypothesis testing theory in a longitudinal microarray study when there are multiple treatments or experimental conditions. The tests developed can be used to detect treatment effects for a large group of genes and discover genes that respond to treatment over time. In the third model, we address a hypothesis testing problem that could arise when array data from different sources are to be integrated. We perform statistical tests by assuming a nested design. In all models, robust test statistics were constructed based on moment methods allowing unbalanced design and arbitrary heteroscedasticity. The limiting distributions were derived under the nonclassical setting when the number of probes is large. The test statistics are not targeted at a single probe. Instead, we are interested in testing for a selected set of probes simultaneously. Simulation studies were carried out to compare the proposed methods with some traditional tests using linear mixed-effects models and generalized estimating equations. Interesting results obtained with the proposed theory in two cancer genomic studies suggest that the new methods are promising for a wide range of biological applications with longitudinal arrays.

Table of Contents

Table of Contents vi

List of Figures ix

List of Tables xi

List of Abbreviations xiii

Acknowledgements xiv

Dedication xv

1 Introduction and motivation 1

2 Non-parametric tests for longitudinal array CGH data 6

2.1 Introduction . . . 6

2.2 Data generation from Affymetrix SNP arrays . . . 13

2.3 Model specification . . . 16

2.4 Testing statistics . . . 18

2.5 Main results based on original observations . . . 20

2.6 Simulation results . . . 34

2.6.1 Simulated data . . . 35

2.6.2 Bootstrap data . . . 40

2.7 A longitudinal study . . . 45

3 Statistical tests for time course microarray data 50

3.1 Introduction . . . 50

3.2 Model specification . . . 55

3.3 Test statistics . . . 57
3.4 Main results based on original observations . . . 62
3.5 Simulation results . . . 89
3.5.1 Type I error rate analysis with simulated data . . . 90
3.5.2 Power analysis with bootstrap data and simulated data . . . 96
3.6 Real data analysis . . . 106

4 Rank-based Hypothesis Testing in Unbalanced Heteroscedastic Nested Design 113

4.1 Model specification . . . 113
4.2 Test statistics . . . 114
4.3 Main results based on original observations . . . 115
4.4 Main results based on ranks . . . 118

5 Summary and future studies 125

5.1 Summary of the current study . . . 125
5.2 Future studies . . . 126
5.2.1 Spatially correlated image data . . . 126
5.2.2 Genetic interaction and gene networking . . . 127
5.2.3 High dimensional data integration . . . 127

Bibliography 139

A R codes for data analysis 140

A.1 R functions for longitudinal aCGH study . . . 140
A.1.1 R function for the test of the probe effect by NPT . . . 141
A.1.2 R function for the test of the time effect by NPT . . . 143
A.1.3 R function for the test of the probe and time interaction by NPT . . . 144
A.1.4 Sample codes for LME and GEE calculations . . . 148
A.2 R functions for longitudinal microarray study with treatment groups . . . 150
A.2.1 R function for the test of the treatment effect by NPT . . . 150

A.2.2 R function for the test of the time effect by NPT 152
A.2.3 R function for the test of the gene effect by NPT 154
A.2.4 R function for the test of the treatment and time interaction by NPT 156
A.2.5 R function for the test of the treatment and gene interaction by NPT 158
A.2.6 R function for the test of the gene and time interaction by NPT . . . 161
A.2.7 Sample codes for LME and GEE calculations 165

List of Figures

2.1 *Demography of human chromosome in normal and tumor cells. The left panel shows the normal complete genome for a male, 22 pairs of autosomal chromosomes plus 2 sex chromosomes, X and Y. Different chromosomes are drawn with distinct colors. The right panel shows the genome from a breast cancer cell line. The colors display the rearrangement of chromosome segments from the normal cell to the tumor cell.* 7

2.2 *Screenshot of Affymetrix Genotyping Analysis software that generates DNA copy numbers from aCGH intensity data. GSACN are Gaussian smoothed copy numbers that are used for further data analysis.* 9

2.3 *A schematic of generation and labelling of probes for hybridization of Affymetrix SNP arrays.* 15

2.4 *The plots of DNA copy numbers in chromosome 7q of normal and glioma samples. Red plots denote the copy numbers of glioma SNPs, and blue plots denote the copy numbers of normal SNPs. The x axis showed the genomic positions of each SNP on chromosome 7q.* . 42

2.5 *The power curves of balanced design with an AR(1) correlation.* 43

2.6 *The power curves of unbalanced design with unstructured correlation.* 43

2.7 *The power curves of unbalanced design with unstructured correlation* 44

2.8 *The power curves of unbalanced design with unstructured correlation.* 45

3.1 *The power curves of testing the treatment effect for unbalanced design with unstructured correlation correlation. There were 2 time points in the experiment.* 99

3.2 *The power curves of testing the treatment effect with 0.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 100

3.3 *The power curves of testing time effect with up to 2% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 101

3.4 *The power curves of testing time effect with up to 1% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 102

3.5 *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 102

3.6 *The power curves of testing time effect with up to 2% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 103

3.7 *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 104

3.8 *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 105

3.9 *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.* 105

3.10 *The histogram showed the distribution of the size of the 548 gene sets used for data analysis.* 108

3.11 *The plots of log odds ratio for the first 10 selected gene sets.* 110

3.12 *The plots of log odds ratio for the second 10 selected gene sets.* 111

List of Tables

2.1 Estimated type I error estimate of the test of no probe effect at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. There are 5 replications in the design. 37

2.2 Estimated type I error of the test of no probe effect at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. The number of time points is 2. For each experiment, Four fifths of probes have 4 replications, and the remaining one fifth of probes have 6 replications. 38

2.3 Estimated type I error of the test of no probe effect at 0.05 level. The data from the same probe follow unstructured correlation. The number of time points is 5. For each experiment, Four fifths of probes have 4 replications, and the remaining one fifth of probes have 6 replications. 38

2.4 Estimated type I error estimate of the test of no time effect at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. For each probe, the number of replications is either 4 or 6. 39

2.5 Estimated type I error of the test of no time effect at 0.05 level. The data from the same probe follow unstructured correlation. For each simulation, there are 8 time points. For each probe, the number of replications is either 4 or 6. 40

2.6 Estimated type I error rates of the test of no interaction of probe and time effects at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. For each probe, the number of replications is either 4 or 6. . . . 40

2.7 Estimated type I error rates of the test of no interaction of probe and time effects at 0.05 level. The data from the same probe follow unstructured correlation with correlation =0.5. For each probe, the number of replications is either 4 or 6. 41

2.8 Summary of significant P values (< 0.05) calculated by NPT methods for each chromosome arm. 47

2.9 Summary of the copy number alterations detected for both primary and relapse tumors. 49

3.1 Estimated type I error estimate of the test of no treatment effect at 0.05 level in unbalanced design. The data from the same gene follows AR(1) with correlation =0.5. 92

3.2 Estimated type I error estimate of the test of no treatment effect at 0.05 level in unbalanced design. The data from the same gene have unstructured correlation. 93

3.3 Estimated type I error of the test of no time effect at 0.05 level. The data from the same gene followed unstructured correlation. For each simulation, there are 2 time points. 94

3.4 Estimated type I error rates of the test of no gene effect at 0.05 level. The data from the same gene follow unstructured correlation. There were either 2 or 5 time points for repeated measures. 95

3.5 Estimated type I error of the test of no treatment*time interaction at 0.05 level. The data from the same gene followed unstructured correlation. For each simulation, there are 2 time points. 96

3.6 Estimated type I error of the test of no treatment*gene interaction at 0.05 level. The data from the same gene followed unstructured correlation. For each simulation, there are 2 time points. 96

3.7 Estimated type I error rates of the test of no time*gene effect at 0.05 level. The data from the same gene follow unstructured correlation. There were either 2 or 5 time points for repeated measures. 97

3.8 The IL-2 regulated gene sets. 109

List of Abbreviations

aCGH	array comparative genome hybridization
AR(1)	a first-order autoregressive process
BAC	bacteria artificial chromosome
CBS	circular binary segmentation
CGH	comparative genome hybridization
ChIP-chip	chromatin immunoprecipitation chip
EM	expectation-maximization
FDA	US Food and Drug Administration
FDR	false discovery rate
FISH	fluorescent in situ hybridization
GEE	generalized estimating equations
GIS	geographic information systems
HMM	hidden Markov model
LASSO	least absolute shrinkage and selection operator
LME	linear mixed-effects model
MRI	magnetic resonance imaging
NPT	nonparametric test
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
RT-PCR	real-time polymerase chain reaction
SNP	single nucleotide polymorphism

Acknowledgments

I would like to thank my major advisor, Dr. Haiyan Wang. Without her close guidance, invaluable assistance and advice, and tremendous patience, I cannot reach this point. I am indebted to her in my whole career.

I would like to thank Dr. Paul Nelson and Dr. James Neill for serving on my committee. Their consistent supports and their efforts for reviewing and commenting on my research work are highly appreciated.

I would like to thank Dr. Ruth Welti and Dr. Michael Dryden who agreed to serve on my committee. I appreciate their time and efforts for reviewing and commenting on my work.

I would like to thank all faculty members in Department of Statistics at Kansas State University. I learned a great deal from taking their courses. Their kindness and willingness to help me will never be forgotten.

Finally, I would like to thank my wife, Linglin Xie. Her trust and consistent support have made it possible to complete the research work.

Dedication

This work is dedicated to my wife, Linglin Xie.

Chapter 1

Introduction and motivation

Advances in high through-put technologies have shifted the focus of scientists from mRNA arrays with known genes to DNA arrays that can scan the entire genome of an organism. This also enabled the transition from study of individual genes to examination of regions of a chromosome. One of such techniques is array Comparative Genomic Hybridization (aCGH), in which the whole genome DNA information of an organism can be scanned onto a chip with high resolution. The unit on the chip is referred as a probe. A probe can contain mutations such as a singular nucleotide polymorphism (SNP), or contain a long sequence of DNA such as BAC clone. There may be multiple probes for a single gene. Though the goals of their specific biological investigations may be different, biologists all need statistical tools to identify DNA segments or regions that exhibit differences for the sample of interest compared with a reference sample. Some examples are identification of DNA regions that show DNA copy number variations for disease versus normal sample, searching for enriched DNA fragments using microarray intensity from chromatin immunoprecipitation (ChIP-chip) to find transcription factor (TF)-binding sites.

Currently, many statistical methods have been used to identify the regions of aberration when the signal intensities on different chips are from independent samples. These include hidden Markov models (Fridlyand et al. 2004; Li et al. 2005; Du et al. 2006), change point analysis (Olshen et al. 2004), local smoothing (Hupe et al. 2004; Zheng et al. 2007),

Bayesian maximum a posteriori probabilities (Daruwala et al. 2004), hierarchical clustering (Wang et al. 2005), regression (Reiss et al. 2007), EM algorithm based edge filtering (Myers et al. 2004) and Bayesian hierarchical model (Gottardo et al. 2008). Most of these methods assume independence or Gaussian distribution with piecewise constant variance for the (log ratio of) intensities of the probes. Even though such assumptions in distribution can simplify the modeling, they may be unrealistic for the complex DNA genomic data and therefore the accuracy of these methods in practice can be a problem. For example, Lai et al. (2005) compared several methods and found that hidden Markov models did not detect any of the three amplified regions in a glioma aCGH data even though they detected smaller regions in their simulated data. On the other hand, the EM algorithm based edge filtering method (Myers et al. 2004) found all three regions in the glioma data to be significant but did not detect the presence of aberrations in their simulated data.

Though many methods are available for independent samples, there are rarely any publications providing methods for identification of the event regions when the experiments have longitudinal aCGH arrays. Recently, research with aCGH arrays for multiple time points have been under development to study the genetic basis of cell development and tumor progression. In these studies, the dynamic behavior of genomic DNA is monitored through multiple chips at different cell growth stages. For instance, time course aCGH studies of in vivo lymphoma and leukemia as well as in vitro tumor cell lines are under investigation for progressive DNA copy number changes in Abbott Laboratories. The examples below justify the significance of methods for longitudinal study using aCGH data.

Example. About 15% of children with Wilms tumor suffer from relapse with nearly half dying of their disease despite aggressive second line treatment regimens. To determine the molecular genetic changes associated with the progression or relapse of Wilms tumor, aCGHs were obtained on ten patients at initial diagnosis and recurrence (Natrajan et al. 2007). Using real-time reverse-transcription polymerase chain reaction (RT-PCR), the authors observed the acquisition of a number of additional molecular alterations between the time of

diagnosis and subsequent relapse, such as a small deletion at 14q12 in four cases, loss of the entire X chromosome, and gain of the whole of 15q and chromosome 5. Unfortunately, the paired t-test they applied did not find any significant recurrent changes in copy number after correction for multiple testing. There were, however, acquired alterations which occurred in more than one relapsed tumor including gain of 5p, 8p12, 15q, 16p, and 20q, as well as the loss of 17p between the time of diagnosis and subsequent relapse. This example makes it clear that methods for detection of event regions using longitudinal aCGH arrays remain to be an important topic to be investigated.

Additional examples can be seen from the research of members in the K-State Ecological Genomics Institute who study how organisms respond to changing environments over long and shorter evolutionary time scales. These include microarray experiments done to study the environmental stresses (such as drought and nitrogen changes) on big bluestem tall grass; expression profiling to examine cellular and molecular responses of aquatic organisms to various environmental stressors including pesticides, heavy metals, nutrients, and oxygen depletion; cellular mechanisms of tomato plants in defensive response to a viral pathogen (tomato spotted wilt virus) and an arthropod herbivore (two-spotted spider mite); genetic responses of soil nematode to changes in soil chemistry caused by nitrogen addition and fire, etc. With all these data produced that require matching statistical tools to decipher the information from biological and pathological processes, statistical methodology for analysis of longitudinal arrays needs to be developed.

The temporal component in the examples above is an inherent part of the study for discovery of important genes or transcriptional activity over developmental stages. Repeated measurements over time on the same subject induce correlations that need to taken into account. Due to the high cost of array experiments, a large sample size assumption is usually impractical. This, together with the high dimensionality, makes the likelihood based optimal test procedures unmanageable or not applicable. See Wahba (1990), Brumback and Rice (1998), Fan and Lin (1998), Huang and Lu (2000, 2001), and Wang (2002) for

examples of innovative models for designs with high-dimensional data. While such models have had considerable impact in theory and applications, their applications in genomic data are hindered due to the distributional assumptions that are restrictive or can not be justified with small sample sizes. For example, many of the models in aforementioned references assume normality or require large sample sizes. But it has been established that such data do not follow a normal distribution (Daruwala et al. 2004; Sidorov et al. 2002; Zhao et al. 2004). Therefore, these methods have limited application in genomic studies.

In many aCGH data, the disease sample and reference sample were hybridized on the same chip. Then the ratio or log ratio of the intensities of the disease and reference samples are used as the observations. In this case, no additional fixed factors are necessary in the model. In other cases, additional fixed factors are necessary to account for the effect of multiple experimental conditions. In this thesis, nonparametric models are developed for longitudinal high dimensional data with or without additional fixed factors. Test procedures for the common hypotheses of interest under each of the models based on original observations are then constructed. The asymptotic distributions of the test statistics are obtained under the non-classical setting in which the number of variables is large while the number of replicates is small. Simulation studies were conducted to evaluate the new test procedures. Applications of the new theory on genomic data from cancer studies are presented. The methods in this thesis are based on a general model set up that allows robust inference in presence of temporal correlations for heteroscedastic high dimensional low sample size data. They provide flexible tools for nonparametric hypothesis testing and can be applied by a wide range of scientists to accelerate novel gene discovery such as identification of biomarkers to control tumor progression, important genes in pest and plant/animal/human interactions, crucial genes for plants, animals or human to acquire tolerance to environmental stresses, etc.

The rest of the thesis is organized as follows. Chapter 2 is devoted to nonparametric tests for high dimensional longitudinal arrays when no additional fixed factors are in the model.

We will focus our discussion in the setting of time course aCGH study to detect DNA copy number variations. Chapter 3 gives the testing procedures for longitudinal arrays when multiple experimental conditions exist. Chapter 4 presents the theory for rank tests for nested design in high dimensional data. Summaries and future research topics are described in Chapter 5. A short introduction about Affymetrix SNP array technology and the R code for the tests are given in Appendices.

Chapter 2

Non-parametric tests for longitudinal array CGH data

Tumor cells usually undergo dramatic chromosome changes resulting in gain or loss of DNA copy numbers. High throughput technologies have made it possible to simultaneously examine DNA copy numbers at thousands or millions of sites of a genome. A time course array study enables discovery of DNA copy number variation during tumor development. In this chapter, we present robust new statistical tests for detecting the DNA regions with copy number variation. Simulation studies show that the proposed methods are robust against non-normality and have higher power than linear mixed-effects models (LME) and generalized estimating equations (GEE). The theory is applied to a longitudinal array study with tumor samples collected from Wilms' patients at both diagnosis and relapse.

2.1 Introduction

The complete genomic information is conveyed by twenty-three pairs of chromosomes in normal human tissue. Enormous efforts have been dedicated to investigating the association of DNA copy number alterations with disease. For normal tissues, each DNA segment has two copies. However, tumor cells undergo complicated pathological progression. Their DNA is often subject to translocation, amplification, and deletion, which leads to DNA copy number

abnormality (Fig. 2.1). DNA amplification may cause over-expression of the encoded genes, and DNA deletion may cause under-expression of the genes. Tumor biomarkers have been intensively investigated in both academy and pharmaceutical industry. They are potentially to be used in various stages of clinical management decisions, such as risk assessment, diagnostic testing, prognostic stratification, and chemotherapy selection (Forozan et al. (2000)). Chromosome alterations associated with tumor progression are promising biomarker candidates. With the encouragement of FDA, the DNA copy number signature of a cancer patient is likely to eventually serve as a basis for considering personalized medicine/therapy.

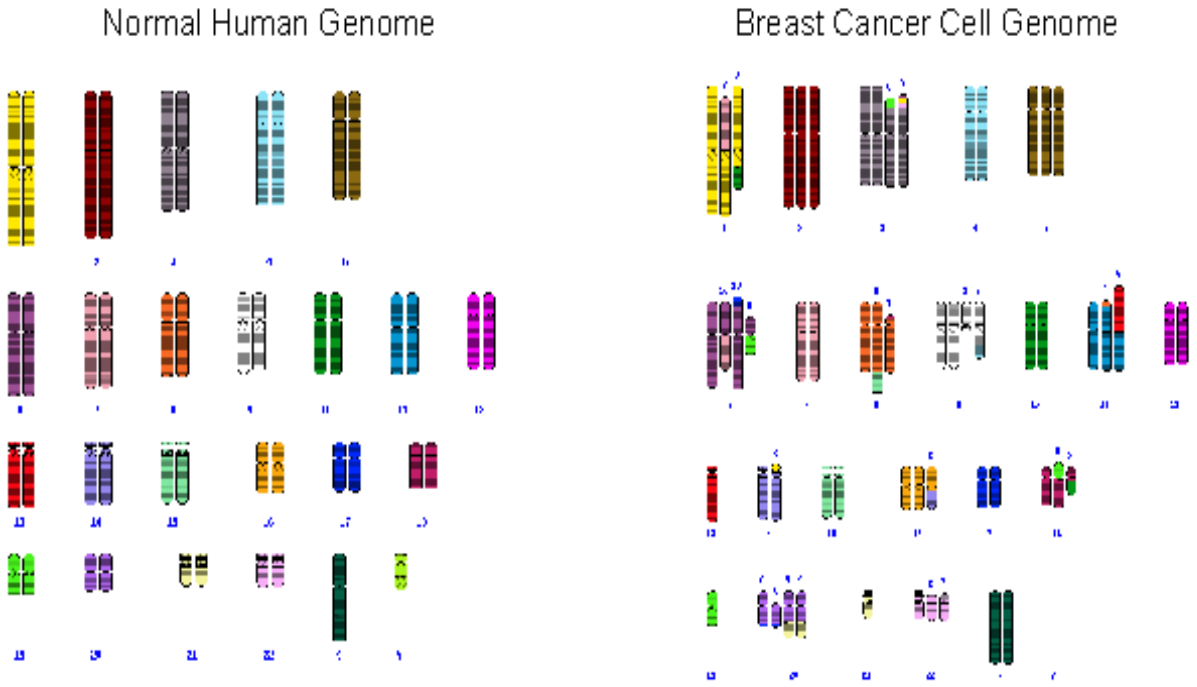
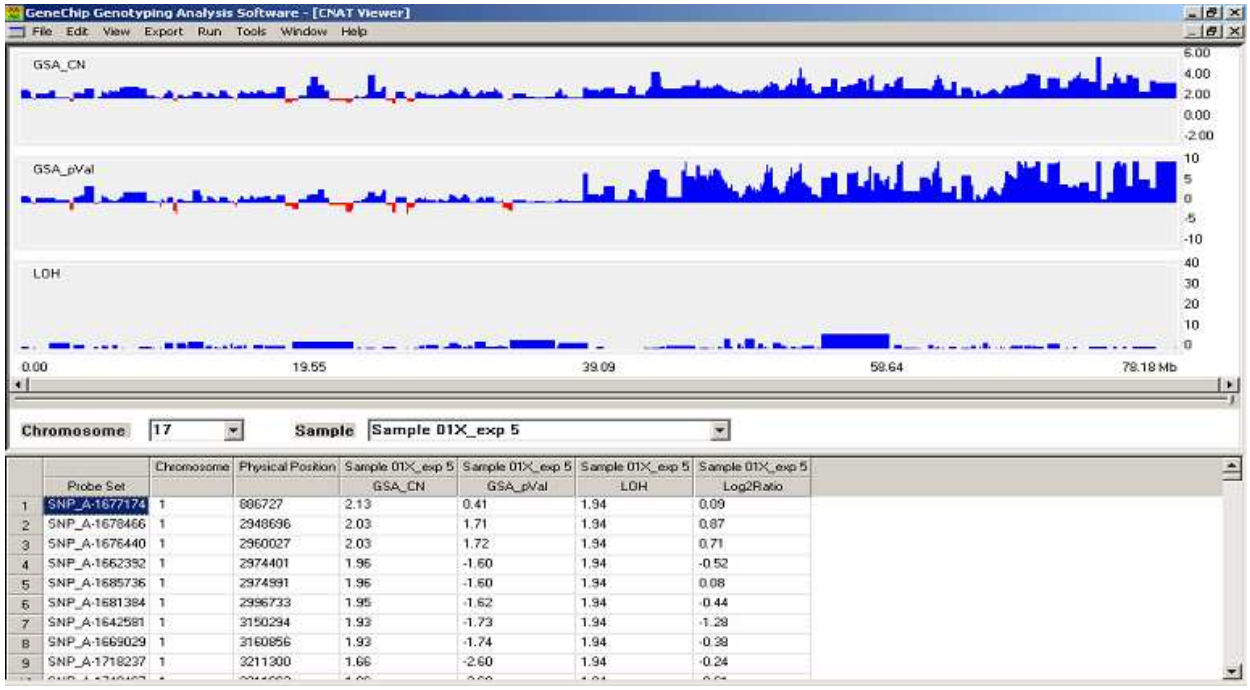


Figure 2.1: *Demography of human chromosome in normal and tumor cells. The left panel shows the normal complete genome for a male, 22 pairs of autosomal chromosomes plus 2 sex chromosomes, X and Y. Different chromosomes are drawn with distinct colors. The right panel shows the genome from a breast cancer cell line. The colors display the rearrangement of chromosome segments from the normal cell to the tumor cell.*

The molecular-cytogenetic method used to detect DNA copy number changes is called comparative genomic hybridization (CGH). Fluorescent in situ hybridization (FISH) and quantitative PCR (qPCR) have been widely used to detect chromosome aberration. However, these traditional techniques are low in resolution, slow in production, and labor inten-

sive. With the rapid advance of high throughput technology, array based technology has been increasingly used in CGH research today (Pinkel and Albertson (2005a)). Array CGH (aCGH) is a microarray-alike technique which detects hundreds of thousands of chromosome sites simultaneously with specific DNA sequences (probes). CGH array detects chromosome DNA copy number, whereas mRNA microarray detects the amount of messenger RNA. In comparison, aCGH reflects genome features, and microarray represents gene expression. Because RNA is a volatile macromolecule which has a very short half life cycle and DNA is relatively stable, aCGH signal is much more stable than microarray. In statistical analysis, aCGH data have less noise and yield more reliable results. Figure 2.2 shows the typical signal of a sequence of aCGH probes. The probe of aCGH is mainly of two types: BAC (Bacterial Artificial Chromosome) clone and DNA oligonucleotide. BAC clones provides the resolution on the order of 1 Mb (Greshock et al. 2004; Pinkel and Albertson 2005b). In the last few years, DNA oligonucleotide arrays have become popular for CGH because they can offer much higher resolution (Brennan et al. 2004). An example of oligonucleotide probe is to use singular nucleotide polymorphism (SNP) to design DNA marker. A chip based on SNP is referred to as a SNP array (Kennedy et al. (2003)). For examples, there are four versions of widely-used SNP arrays manufactured by Affymetrix, each consisting of 100K, 250K, 500K, and 1M SNPs, respectively. Another company, Agilent, provides a 250K SNPs array. Compared to microarray, which usually contain 5,000-50,000 probe sets, aCGH data analysis is more suitable for novel gene discovery but also raises challenge for statistical methodology, computational cost and memory usage. In section 2.2, we give a detailed introduction of copy number generation from Affymetrix SNP array.

In aCGH study, we are often interested in copy number variations for large genomic segments. For example, each chromosome has two arms, p and q, that are connected by a centromere. Chromosome rearrangement often causes one arm to be translocated, duplicated or lost. The copy number changes of a chromosome arm will affect thousands of SNPs located in it. A number of statistical tools have been developed to detect gain or loss of



	Probe Set	Chromosome	Physical Position	Sample 01X_exp 5 GSA_CN	Sample 01X_exp 5 GSA_pVal	Sample 01X_exp 5 LOH	Sample 01X_exp 5 Log2Ratio
1	SNP_A-1677174	1	886727	2.13	0.41	1.94	0.09
2	SNP_A-1678466	1	2948696	2.03	1.71	1.94	0.87
3	SNP_A-1676440	1	2960027	2.03	1.72	1.94	0.71
4	SNP_A-1662392	1	2974401	1.96	-1.60	1.94	-0.52
5	SNP_A-1685736	1	2974991	1.96	-1.60	1.94	0.08
6	SNP_A-1681384	1	2996733	1.95	-1.62	1.94	-0.44
7	SNP_A-1642581	1	3150294	1.93	-1.73	1.94	-1.28
8	SNP_A-1669029	1	3160856	1.93	-1.74	1.94	-0.38
9	SNP_A-1718237	1	3211300	1.66	-2.60	1.94	-0.24

Figure 2.2: *Screenshot of Affymetrix Genotyping Analysis software that generates DNA copy numbers from aCGH intensity data. GSACN are Gaussian smoothed copy numbers that are used for further data analysis.*

DNA segment. Fridlyand et al. (2004) and Zhao et al. (2004) used hidden Markov models (HMM) to estimate probe copy numbers; Olshen et al. (2004) proposed a circular binary segmentation (CBS) algorithm to detect the break points of DNA segments; Daruwala et al. (2004) and Sabatti and Lange (2008) fitted copy number variation with Bayesian models; Hsu et al. (2005) smoothed aCGH signal with wavelet; Zou and Hastie (2005), Reiss et al. (2007), Tibshirani and Wang (2008) selected important genomic regions by constraining the regression parameters with L_1 or L_2 Norm. Most of these methods assume a specific distribution for the (log ratio of) intensities of the probes such as normal, log-normal or Poisson distribution. Despite the large amount of efforts made to justify those distributions (Sidorov et al. 2002; Hoyle et al. 2002), it has been vigorously argued as to whether the fitting of real image data with a well-defined distribution is adequate (Kerr et al. 2000; Konishi 2004). Low convergence rates and slow computations are another problem with some of high-computation oriented techniques. Based on our experiments, it affects the

Bayesian-based models and regression-based models such as LASSO and elastic net. The computational difficulty makes those methods impossible to be applied to dense SNP arrays. Because of high dimensional and noisy nature of aCGH data, many of these methods could not provide reliable and consistent results. Lai et al. (2005) discussed the strengths and limitations of 11 methods. For examples, HMM was sensitive to detect small abnormal regions in simulated data, but failed to detect any of the three amplified regions in a glioma aCGH data; CBS successfully detected all three amplifications of glioma data but blurred the break points.

In a cancer study, one of the central goals is to understand tumor development and progression. Studies are designed to monitor the dynamic behavior of genomic DNA. For instances, chromosomal instability during neoplastic progression was investigated for Barrett's esophagus patients with Affymetrix 100K SNP arrays (Lai et al. 2007), and relapse of Wilms' tumor was investigated with BAC clone CGH array (Natrajan et al. 2007). In these studies, researchers are interested in identifying a genome signature consisting of DNA aberration regions that is associated with a disease outcome and a drug response. As patients are repeatedly measured, a proper within-subject correlation should be considered for any reasonable statistical interpretations. Technically, the within-subject correlation over time is the same as within-cluster correlation such as genomes within a family. An example of within-family correlation can be found in the Framingham Heart Study directed by the National Heart, Lung, and Blood Institute (Kottgen et al. 2008). They studied the genotypes of original and Offspring cardiovascular disease patients with Affymetrix 100K SNP array. The goal of the study is to identify the genetic factors underlying cardiovascular disease and other disorders. Any methodology developed for longitudinal aCGH data should be able to be applied to such within-cluster correlation data.

Most data analyses for an longitudinal microarray are based on statistical tests for individual gene that is adjusted for multiple tests by using false discovery rate (FDR) procedure. Linear mixed-effects models and generalized estimating equations (GEE) are widely used

for longitudinal data analysis (Liang and Zeger (1986); Diggle et al. (2002)). While they have proven to be useful tools for repeated measures with large sample size, the adequacy of model fitting is of concern for high dimensional data analysis (Fan and Zhang (2000)). Park et al. (2003) used a two-stage ANOVA model to calculate the P values for each gene. At the first stage, a time effect is tested. The residuals of time effects are then used in a permutation test. For their analysis, the study need to be balanced and the within-subject correlation cannot be strong to achieve sufficient statistical power. Guo et al. (2003) proposed a modified Wald statistic to test the differential expression of each gene over time. The Wald statistic converges to a χ^2 distribution under the null hypothesis when the number of subjects is sufficiently large. Each gene is assigned a gene-specific score that is calculated by the Wald statistic that accounts for within-subject correlation. A permutation test is then performed to compute the false discovery rate (FDR) for each gene. Storey et al. (2005) used a mixed model with a polynomial mean function to detect significant genes across time points between treatment and control groups. Under the null hypothesis of no differential expression, the two groups are assumed to have the same population average time curve. The population mean curve is modeled for the profile of each gene, and an F statistic is then calculated based on it. P values are adjusted by FDR to determine significantly differentiated genes.

All the above methods are based on univariate test for an individual gene by taking into account of its within-subject correlation. Despite their wide applications in expression microarray studies, their usage for aCGH study is limited because a FDR adjustment is not sensitive enough in detecting small copy number variation, which is often important (Storey and Tibshirani (2003)). Analysis of aCGH data is usually based on segmented probes for chromosome rearrangement affecting a large number of probes located within the genomic region. A rich class of techniques such as HMM and CBS as discussed earlier have been proposed to segment DNA for independent samples. For dependent aCGH samples, researchers are currently short of robust techniques. Tsai and Qu (2008) performed hypothesis testing

for a set of genes by applying a non-parametric time-varying coefficient model. The within-subject correlation was taken into account by a quadratic inference function (QIF). A QIF derived from GEE is asymptotically χ^2 distributed when the number of replications goes to ∞.

Due to the high cost of array experiments, a large sample size is usually not desired. Therefore, methods based on large sample size have limited application in aCGH study. In addition, the computational cost is substantially higher for time course analysis than for the static experiments. The goal of this chapter is to provide a series of test statistics for detect copy number variation of a DNA segment in a longitudinal aCGH study. The test statistics should be robust to non-normality. They can also be used for other high dimensional low replicated data with within-subject correlation. The proposed test statistics are applied to unbalanced designs and heteroscedastic covariance structures as well. The method can be used to identify genomic signatures with a test-based clustering algorithm.

In aCGH study, DNA copy number can be inferred by $\log_2$ ratio of the disease and reference samples when both samples are hybridized onto the same chip. A positive $\log_2$ ratio indicates gain in copy number and a negative value indicates loss in copy number. Due to experimental and biological factors such as purity of a sample, the $\log_2$ ratio does not appear as the magnitude of copy numbers. For example, a frequent phenomenon in the analysis of primary tumors is normal cell contamination caused by imperfect dissection. Generally, pathologists make sure that each tumor sample contains no more than 50% (or 30%) of normal cells. The purity of the tumor sample increases as the contamination proportion decreases. Another factor that affects the copy number estimation is that not all tumor cells may have acquired a given aberration. These factors make estimation of a true copy number impossible. Here we aim at testing whether DNA sections within a DNA region have common copy numbers. The test then can be used to partition the chromosome into sets of the same copy number segment.

The outline of this chapter is as follows. Section 2.2 presents the processes of copy number generation illustrated with Affymetrix 100K SNP array. In section 2.3, we describe the study design and the model specification. Test statistics are provided in section 2.4. Details of asymptotic theories for original observations and the corresponding proofs are provided in section 2.5. Simulation study is presented in section 2.6. Type I error rates were estimated under various distributions in simulations, and power analysis was compared to LME and GEE with bootstrap data. In section 2.7, we apply our methods to a Wilms' tumor study.

2.2 Data generation from Affymetrix SNP arrays

In this section, we give an introduction of the array CGH technology. Array designs vary manufacturer to manufacturer, and sometimes even vary version-to-version of the same manufuacturer. For simplicity, we base our discussion on Affymetrix 100K SNP array.

SNPs are sequence changes that arose once during evolution. Public efforts have so far identified over two million common human SNPs. Affymetrix designed the 100K SNP array consisting of more than 110K SNPs distributed across a human genome. The SNP arrays can be used for genetic linkage analyses, genotyping calling, and DNA copy number variation study (Kennedy et al. 2003).

Typically, only two of the four possible bases at an SNP are present in human. If we denote the two alleles at an SNP by the letters a and b, then each person has one of three possible genotypes a/a, a/b, or b/b.

In Affymetrix SNP array, each SNP is assessed by 40 probes, each 25 bases long. Of the 40 probes, 20 are match probes that perfectly hybridize with one of the two alleles, and 20 are mismatch probes intended to measure the level of cross-hybridization. Among the 20 match probes, 10 probes are complementary to allele a, and 10 probes are complementary to allele b. Each set of 10 match probes is further subdivided into two subsets of 5 probes; one

subset is complementary to the sense strand and the other subset to the antisense strand of the DNA molecule. This leads to four probe subsets: sense (s) a, antisense (t) a, sense b, and antisense b. Each mismatch probe subset is paired with one subset of match probes and differs from it at the base in the central position of the oligonucleotide.

The sample genomic DNA is processed with the following steps: (1) The DNA is broken into small fragments by restriction enzyme digestion, (2) the fragments are amplified by the polymerase chain reaction (PCR), (3) fragment copies are labeled with dye molecules to distinguish the two alleles, (4) labeled fragment copies are hybridized with the array, and (5) the intensity of the fluorescent signal at each spot is measured. The DNA labeling process is illustrated in Figure 2.3.

Numerous algorithms have been proposed to summarize and to analyze the raw fluorescent intensity data derived from labeling of the array (Yang et al. 2002; Irizarry et al. 2003). Here we focus our discussion on the copy number algorithm recommended by Affymetrix (Affymetrix 2006). The intensity data are first subjeted to probe/SNP filtering for quality control. Users can exclude mismatch probes for subsequent analysis. Additionally, users can exclude SNPs based on the length of the PCR fragment with which they hybridize. It have been shown that the exclusion of SNPs on larger PCR fragment sizes improve analytical accuracy. Secondly, probe intensities are normalized across multichips with the goal to reduce experimental noise due to chip-to-chip variation, background, and relative variation in the performance of probes interrogating a given SNP. Various methods are available for data normalization. For examples, median scaling, quantile normalization, and Gaussian smoothing are a few widely used approaches (Quackenbush 2002).

If we use both perfectly matched and mismatched probes in the analysis, we need to summarize the relative measure with a discrimination score (D). For the ith allele, let PM_i denote the average normalized intensity of the perfectly matched probes, and MM_i for the average of the mismatched probes. The discrimination score for the ith allele is

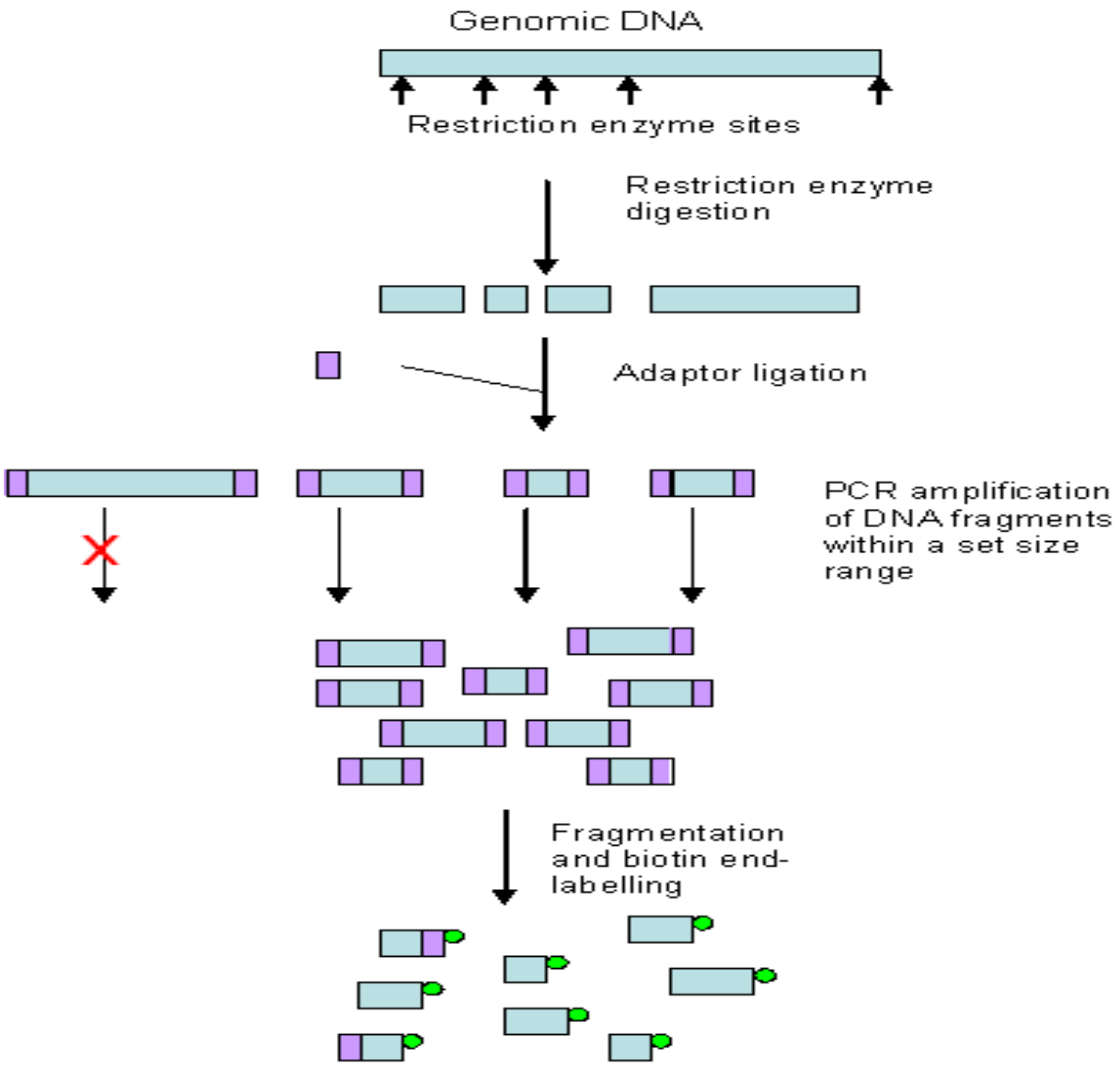


Figure 2.3: *A schematic of generation and labelling of probes for hybridization of Affymetrix SNP arrays.*

$$D_i = \frac{PM_i - MM_i}{PM_i + MM_i}.$$

The raw copy number (CN) estimation is based on the log2ratio between tumor sample and a reference sample. In practice, we use 48 female normal samples provided by Affymetrix as global references. The copy number is generated for every allele. λ_i is the raw CN for the ith allele based on sample score S_i and reference score R_i.

$$\lambda_i = log_2 \frac{S_i}{R_i},$$

where S_i and R_i are the average intensities of perfectly matched probes if only perfectly matched probes are used, or are discrimination scores if mismatched probes are included.

Then Total Copy Number (TCN) of SNP k can be estimated with *sumLog* formula by summing both alleles.

$$TCN_k = log_2 \frac{S_{kA}}{R_{kA}} + log_2 \frac{S_{kB}}{R_{kB}},$$

where

R_{kA}, R_{kB}: scores in the allele a, b for SNP k of the reference;

S_{kA}, S_{kB}: scores in the allele a, b for SNP k of the tumor sample.

Therefore, for each SNP k, three raw CN esimates are generated: TCN, CN for allele a (λ_{kA}), and CN for allele b (λ_{kB}). For DNA copy number analysis discussed in this thesis, TCN are the input data for our statistics. TCN are continuous numerical data. They can be transformed to integer copy number using hidden Markov models (HMM) or mixture Gaussian models (Hodgson et al. 2001; Fridlyand et al. 2004). We prefer using raw TCN data to avoid possible systemtic errors incurred in the transformation.

2.3 Model specification

In this section, we consider statistical analysis of high dimensional data with each subject repeatedly measured over time. We will focus our discussion on analyses applied to a time course aCGH study.

Let X_{ijk} be the jth measurement of the ith probe from subject k ($i = 1, ..., I$; $j = 1, ..., J$; $k = 1, ..., n_i$). The number of probes is large, whereas the number of time

points and the number of subjects are fixed. The design is assumed to be either balanced or unbalanced, in that the number of subjects may vary for different probes. For example, if all aCGH data come from a same version of a SNP array, the design will be balanced. However, to take advantage of limited sources, we often need to import aCGH data from different sources. For example, if data from Affymetrix 100K and 250K arrays are to be combined, the design will be unbalanced. In this case, the SNPs shared between 100K and 250K arrays have more samples than the SNPs that only exist in 250K arrays. The probe copy numbers are modeled by

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{2.3.1}$$

where $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$, μ is the overall mean, α_i represents the effect of the ith probe, β_j represents the effect of the jth time point, and γ_{ij} represents the interaction effect of probe and time. The error terms ε_{ijk} have mean 0, and they are correlated for repeated measures from the same subject and the same probe. In other words, ε_{ijk} and $\varepsilon_{i'j'k'}$ are independent if $i \neq i'$ or $k \neq k'$, and they are only dependent when i=i', $j \neq j'$, and k=k'. Note that normality is not assumed for ε_{ijk}.

The dependence of repeated measurements within an individual was taken into account in the model fitting procedure. The within-subject correlation structure is not necessarily homogeneous, it could vary for different probes. The covariance of a probe over time is possibly dependent on its copy number or its location in a chromosome. Furthermore, experiments of biological time course study are often not evenly spaced in time. Therefore, the same correlation structure may not be appropriate. We apply a heteroscedastic variance structure to the model specified by $Cov(\varepsilon_{ijk}, \varepsilon_{ij'k}) = \sigma_{i,jj'}$.

The tests can be written in terms of the parameters in the model. For aCGH study, the null hypothesis of no copy number variation is equivalent to restricting all α_i to be zero. The test will applied to detect the local DNA copy number changes in a given genome region. If the goal is to find the DNA segments whose copy numbers are altered by time points, we

will test that all β_j equal to zero. The null hypothesis of DNA copy number is the same over all time points is equivalent to restricting all γ_{ij} to zero. It will be used to detect the genome regions that have the same time response.

At the end of the section, we present a summary of notations which will be used in the rest of the manuscript.

$$\widetilde{X}_{i\cdot\cdot} = J^{-1}\sum_{j=1}^{J}\overline{X}_{ij\cdot}, \quad \widetilde{X}_{\cdot j\cdot} = I^{-1}\sum_{i=1}^{I}\overline{X}_{ij\cdot},$$

$$\sigma_{i,jj_1} = Cov(X_{ijk}, X_{ij_1k}) \ for \ any \ k, \ (note \ \sigma_{i,jj} = Var(X_{ijk}) = \sigma^2_{ij}),$$

$$\sigma_{i,jj_1,j_2j_3} = Cov(X_{ijk}X_{ij_1k}, X_{ij_2k}X_{ij_3k}), \ (\sigma_{i,jj_1,jj_1} = \sigma^2_{i,jj_1}).$$

2.4 Testing statistics

In this section, we will use a few modified Wald test statistics and modified F test statistics to provide robust tests for main effects and interactions.

First, we consider to explore whether copy number variation exists in a given genomic region. Statistically, it is a test of no main effect of probe. The test will be very useful in DNA segmentation, by which we want to partition the whole genome into amplified, deleted, and normal regions. Under the null hypothesis, there is no copy number difference within the DNA segment of interest. The null hypothesis is

$$H_0(A): \ all \ \alpha_i = 0, \ for \ i = 1, ..., I.$$

where I is the total number of probes located in this DNA segment.

To test $H_0(A)$, we modified the F statistic used in mixed ANOVA model.

$$F_X(A) = \frac{MST_A}{MSE_A}, \quad MST_A = \frac{1}{I-1}\sum_{i=1}^{I}\sum_{j=1}^{J}(\widetilde{X}_{i\cdot\cdot} - \widetilde{X}_{\cdots})^2, \tag{2.4.1}$$

$$MSE_A = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_i(n_i-1)}\sum_{k=1}^{n_i}(X_{ijk} - \overline{X}_{ij\cdot})(X_{ij_1k} - \overline{X}_{ij_1\cdot}). \tag{2.4.2}$$

where $\widetilde{X}_{i\cdot\cdot}$ is the sample average of $\overline{X}_{ij\cdot}$ for $j = 1, ..., J$, and $\widetilde{X}_{\cdot\cdot\cdot}$ is the sample average of $\widetilde{X}_{i\cdot\cdot}$ for all i. The definition of MST_A slightly different from that of ANOVA in that unweighted averages are used instead of weighted averages. The definition of MSE_A is different from that of the traditional MSE in that the within-subject correlation over time is taken into account.

Secondly, we want to carry out a statistical test of time effect. In a longitudinal aCGH study, researchers are often interested in identifying DNA segments whose copy number varies over time. These are potential genomic signatures indicating tumor progression or regression. The test of time effect is targeted on all probes in a selected genomic region.

The null hypothesis of no time effect is

$$H_0(B): \ all \ \beta_j = 0, \ for \ j = 1, ..., J.$$

In order to test $H_0(B)$ of time effect, we also consider a more general hypothesis $H_0(B_G)$: $L\boldsymbol{\beta} = \mathbf{0}$ where L is a $p \times J$ contrast matrix with full row rank, $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)'$, and $\mathbf{0}$ is a p dimensional zero vector. A modified Wald-type test statistic is used for testing $H_0(B_G)$.

$$W_B = D_B' L' (L\hat{V}_B L')^{-1} L D_B \tag{2.4.3}$$

where $D_B = (\widetilde{X}_{\cdot 1\cdot}, ..., \widetilde{X}_{\cdot J\cdot})'$, and $\hat{V}_B$ is the estimated $J \times J$ covariance matrix for D_B, with the value at the jth row and the j'th column be $\hat{V}_{B,jj'} = I^{-2} \sum_{i=1}^{I} (n_i(n_i - 1))^{-1} \sum_{k=1}^{n_i} (X_{ijk} - \overline{X}_{ij\cdot})(X_{ij'k} - \overline{X}_{ij'\cdot})$.

Finally, the high throughput time course study is often targeted to identify the variables that show time response, such as genes regulated by cell cycle, or genomic regions that reflect progressive tumors. Note that DNA copy number variation implies the changes of expression level for the encoded genes by the DNA segment. The hypothesis test of interaction between probe and time will allow us to identify the candidate gene in the genome region where the probe is located, and whose expression changes over time. The null hypothesis is

$$H_0(AB): \ all \ (\alpha\beta)_{ij} = 0, \ for \ i = 1, ..., I, \ and \ j = 1, ..., J.$$

Similar to the test for main effect of gene, the test statistic for no interaction is also based on a modified F statistic

$$F_X(AB) = \frac{MST_{AB}}{MSE_{AB}}, \tag{2.4.4}$$

where

$$MST_{AB} = \frac{1}{(I-1)(J-1)} \sum_{i,j} (\overline{X}_{ij\cdot} - \widetilde{X}_{i\cdot\cdot} - \widetilde{X}_{\cdot j\cdot} + \widetilde{X}_{\cdots})^2, \tag{2.4.5}$$

$$\begin{aligned} MSE_{AB} &= \frac{1}{I(J-1)} \sum_{i=1}^{I} \sum_{j}^{J} \frac{1}{n_i(n_i-1)} \sum_{k=1}^{n_i} (X_{ijk} - \overline{X}_{ij\cdot})^2 - \\ &\quad \frac{1}{IJ(J-1)} \sum_{i=1}^{I} \sum_{j,j_1}^{J} \frac{1}{n_i(n_i-1)} \sum_{k=1}^{n_i} (X_{ijk} - \overline{X}_{ij\cdot})(X_{ij_1k} - \overline{X}_{ij_1\cdot}). \end{aligned}$$

The asymptotic distribution for each of the test statistics in (2.4.1), (2.4.3), (2.4.4) will be derived in the following sections.

2.5 Main results based on original observations

This section is devoted to developing the asymptotic distribution of the test statistics which are defined in the last section. For simplicity, we use notation $e_{ijk} = X_{ijk} - E[X_{ijk}]$ in the proof.

Theorem 2.5.1. *For testing $H_0(A)$: all $\alpha_i = 0$, let $F_X(A)$ be the statistic given in (2.4.1). If X_{ijk} has finite fourth central moment, then under $H_0(A)$,*

$$\frac{\sqrt{I}(F_X(A) - 1)}{V_A} \xrightarrow{d} N(0,1), \quad \text{as } I \to \infty.$$

where

$$V_A = \sqrt{\tau_A}/\sigma_A, \tag{2.5.1}$$

with

$$\tau_A = \frac{1}{IJ^2}\sum_{i=1}^{I}\frac{2}{n_i(n_i-1)}\sum_{j,j_1,j_2,j_3}^{J}\sigma_{i,jj_1}\sigma_{i,j_2j_3},\ \textit{and}\ \sigma_A = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{\sigma_{i,jj_1}}{n_i}.$$

Lemma 2.5.2. *Under the settings and assumptions of Theorem 2.5.1,*

$$MSE_A - \sigma_A \xrightarrow{p} 0 \ \textit{as}\ I \to \infty.$$

Proof:

For any $j, j_1 = 1, ..., J$, note that

$$\begin{aligned}
& E\left[(X_{ijk} - \overline{X}_{ij.})(X_{ij_1k} - \overline{X}_{ij_1.})\right] \\
= \ & E\left[(e_{ijk} - \overline{e}_{ij.})(e_{ij_1k} - \overline{e}_{ij_1.})\right] \\
= \ & E(e_{ijk}e_{ij_1k}) - E(\overline{e}_{ij.}e_{ij_1k}) - E(e_{ijk}\overline{e}_{ij_1.}) + E(\overline{e}_{ij.}\overline{e}_{ij_1.}) \\
= \ & \sigma_{i,jj_1} - \frac{1}{n_i}\sigma_{i,jj_1} - \frac{1}{n_i}\sigma_{i,jj_1} + \frac{1}{n_i^2}\sum_{k,k_1}^{n_i}E(e_{ijk}e_{ij_1k_1}) \\
= \ & \frac{n_i - 1}{n_i}\sigma_{i,jj_1}.
\end{aligned}$$

Thus, we have

$$E(MSE_A) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_i(n_i-1)}\sum_{k}^{n_i}E\left[(X_{ijk} - \overline{X}_{ij.})(X_{ij_1k} - \overline{X}_{ij_1.})\right] = \sigma_A.$$

The result will follow if we show $Var(MSE_A) \to 0$ as $I \to \infty$.

$$\begin{aligned}
|Var(MSE_A)| \ = \ & \frac{1}{I^2J^2}\sum_{i=1}^{I}\sum_{k=1}^{n_i}\frac{1}{n_i^2(n_i-1)^2}\Bigg|Cov[\sum_{j,j_1}^{J}(e_{ijk} - \overline{e}_{ij.})(e_{ij_1k} - \overline{e}_{ij_1.}), \\
& \sum_{j_2,j_3}^{J}(e_{ij_2k} - \overline{e}_{ij_2.})(e_{ij_3k} - \overline{e}_{ij_3.})]\Bigg|.
\end{aligned}$$

Note that

$$
\begin{aligned}
& E\left[(X_{ijk}-\overline{X}_{ij.})(X_{ij_1k}-\overline{X}_{ij_1.})\right]^2 \\
= \;& E\left[(e_{ijk}-\overline{e}_{ij.})(e_{ij_1k}-\overline{e}_{ij_1.})\right]^2 \\
= \;& E[e_{ijk}e_{ij_1k}-\overline{e}_{ij.}e_{ij_1k}-e_{ijk}\overline{e}_{ij_1.}+\overline{e}_{ij.}\overline{e}_{ij_1.}]^2 \\
\leq \;& 4[E(e_{ijk}e_{ij_1k})^2+E(\overline{e}_{ij.}e_{ij_1k})^2+E(e_{ijk}\overline{e}_{ij_1.})^2+E(\overline{e}_{ij.}\overline{e}_{ij_1.})^2] \\
= \;& 4\left[E(e_{ijk}^2e_{ij_1k}^2)+\frac{1}{n_i^2}E(e_{ijk}^2e_{ij_1k}^2)+\frac{1}{n_i^2}E(e_{ijk}^2e_{ij_1k}^2)+\frac{1}{n_i^4}\sum_{k=1}^{n_i}E(e_{ijk}^2e_{ij_1k}^2)\right] \\
= \;& \frac{4(n_i^3+2n_i+1)}{n_i^3}Cov(e_{ijk}^2,e_{ij_1k}^2) \\
< \;& \infty,
\end{aligned}
$$

where the last inequality holds because X_{ijk} has the finite fourth central moment. The first inequality follows from Hölder's inequality,

$$
\left|\sum_{i=1}^{m}z_i\right|^p \leq m^{p-1}\sum_{i=1}^{m}|z_i|^p,\ m\geq 1,\ p>1. \tag{2.5.2}
$$

We have

$$
\begin{aligned}
& \left|Cov[\sum_{j,j_1}^{J}(e_{ijk}-\overline{e}_{ij.})(e_{ij_1k}-\overline{e}_{ij_1.}),\sum_{j_2,j_3}^{J}(e_{ij_2k}-\overline{e}_{ij_2.})(e_{ij_3k}-\overline{e}_{ij_3.})]\right| \qquad (2.5.3) \\
\leq \;& \left|Var[\sum_{j,j_1}^{J}(e_{ijk}-\overline{e}_{ij.})(e_{ij_1k}-\overline{e}_{ij_1.})]Var[\sum_{j_2,j_3}^{J}(e_{ij_2k}-\overline{e}_{ij_2.})(e_{ij_3k}-\overline{e}_{ij_3.})]\right|^{\frac{1}{2}} \\
\leq \;& \left|E[\sum_{j,j_1}^{J}(e_{ijk}-\overline{e}_{ij.})(e_{ij_1k}-\overline{e}_{ij_1.})]^2\right|^{\frac{1}{2}}\left|E[\sum_{j_2,j_3}^{J}(e_{ij_2k}-\overline{e}_{ij_2.})(e_{ij_3k}-\overline{e}_{ij_3.})]^2\right|^{\frac{1}{2}} \\
\leq \;& \left|J^2\sum_{j,j_1}^{J}E[(e_{ijk}-\overline{e}_{ij.})(e_{ij_1k}-\overline{e}_{ij_1.})]^2\right|^{\frac{1}{2}}\left|J^2\sum_{j_2,j_3}^{J}E[(e_{ij_2k}-\overline{e}_{ij_2.})(e_{ij_3k}-\overline{e}_{ij_3.})]^2\right|^{\frac{1}{2}} \\
= \;& J^2\left|\sum_{j,j_1}^{J}E[(e_{ijk}-\overline{e}_{ij.})(e_{ij_1k}-\overline{e}_{ij_1.})]^2\right|^{\frac{1}{2}}\left|\sum_{j_2,j_3}^{J}E[(e_{ij_2k}-\overline{e}_{ij_2.})(e_{ij_3k}-\overline{e}_{ij_3.})]^2\right|^{\frac{1}{2}} \\
< \;& \infty,
\end{aligned}
$$

where the inequalities follow from the properties of variance and Hölder's inequality, and the last equality holds for the finiteness showed previously. Therefore, $Var(MSE_A) = O(I^{-1})$ and $MSE_A - \sigma_A^2 \xrightarrow{p} 0$ as $I \to \infty$.

Lemma 2.5.3. *Under the settings and assumptions of Theorem 2.5.1 and under $H_0(A)$, we have*

$$\sqrt{I}(MST_A - P_A(e)) \xrightarrow{p} 0 \text{ as } I \to \infty,$$

where $P_A(e) = \frac{J}{I}\sum_{i=1}^{I} \tilde{e}_{i..}^2$.

Proof:

Note that under $H_0(A)$,

$$\begin{aligned} MST_A &= \frac{1}{I-1}\sum_{i=1}^{I}\sum_{j=1}^{J}(\tilde{e}_{i..} - \tilde{e}_{...})^2 \\ &= \frac{J}{I-1}\sum_{i=1}^{I}(\tilde{e}_{i..} - \tilde{e}_{...})^2 \\ &= \frac{J}{I-1}\left(\frac{I-1}{I}\sum_{i=1}^{I}\tilde{e}_{i..}^2 - \frac{1}{I}\sum_{i\neq i'}^{I}\tilde{e}_{i..}\tilde{e}_{i'..}\right) \\ &= \frac{J}{I}\sum_{i=1}^{I}\tilde{e}_{i..}^2 - \frac{J}{I(I-1)}\sum_{i\neq i'}^{I}\tilde{e}_{i..}\tilde{e}_{i'..}. \end{aligned}$$

Thus, we have

$$E[\sqrt{I}(MST_A - P_A(e))] = \frac{\sqrt{I}J}{I(I-1)}\sum_{i\neq i'}^{I} E\left[\tilde{e}_{i..}\tilde{e}_{i'..}\right] = 0,$$

and

$$\begin{aligned}
E[\sqrt{I}(MST_A - P_A(e))]^2 &= \frac{IJ^2}{I^2(I-1)^2} E\left[\sum_{i\neq i'}^{I} \tilde{e}_{i..}\tilde{e}_{i'..}\right]^2 \\
&= \frac{J^2}{I(I-1)^2} E\left[\sum_{i\neq i_1, i_2\neq i_3}^{I} \tilde{e}^2_{i..}\tilde{e}^2_{i_1..}\tilde{e}^2_{i_2..}\tilde{e}^2_{i_3..}\right] \\
&= \frac{2J^2}{I(I-1)^2} E\left[\sum_{i\neq i_1}^{I} \tilde{e}^2_{i..}\tilde{e}^2_{i_1..}\right] \\
&= \frac{2J^2}{I(I-1)^2} \sum_{i\neq i_1}^{I} E[\tilde{e}^2_{i..}]E[\tilde{e}^2_{i_1..}] \\
&= O(I^{-1}).
\end{aligned}$$

Therefore, under $H_0(A)$, $\sqrt{I}(MST_A - P_A(e)) \xrightarrow{p} 0$ as $I \to \infty$.

Proof of Theorem 2.5.1: By Lemmas 2.5.2 and 2.5.3, we need only to consider the asymptotic distribution of $Q_A(e) = \sqrt{I}(P_A(e) - MSE_A)$ under $H_0(A)$, where $P_A(e) = \frac{J}{I}\sum_{i=1}^{I} \tilde{e}^2_{i..}$.

We can write

$$\begin{aligned}
Q_A(e) &= \sqrt{I}\left[\frac{J}{I}\sum_{i=1}^{I}\widetilde{e}_{i\cdot\cdot}^2 - \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_i(n_i-1)}\sum_{k=1}^{n_i}(X_{ijk}-\overline{X}_{ij\cdot})(X_{ij_1k}-\overline{X}_{ij_1\cdot})\right] \\
&= \frac{1}{J\sqrt{I}}\sum_{i=1}^{I}\left[(\sum_{j}^{J}\overline{e}_{ij\cdot})^2 - \sum_{j,j_1}^{J}\frac{1}{n_i(n_i-1)}\sum_{k=1}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})(e_{ij_1k}-\overline{e}_{ij_1\cdot})\right] \\
&= \frac{1}{J\sqrt{I}}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\left[\overline{e}_{ij\cdot}\overline{e}_{ij_1\cdot} - \frac{1}{n_i(n_i-1)}\sum_{k=1}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})(e_{ij_1k}-\overline{e}_{ij_1\cdot})\right] \\
&= \frac{1}{J\sqrt{I}}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\left[\frac{n_i-2}{n_i-1}\overline{e}_{ij\cdot}\overline{e}_{ij_1\cdot} - \frac{1}{n_i(n_i-1)}\sum_{k=1}^{n_i}(e_{ijk}e_{ij_1k} - e_{ij_1k}\overline{e}_{ij\cdot} - e_{ijk}\overline{e}_{ij_1\cdot})\right] \\
&= \frac{1}{J\sqrt{I}}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\left[\frac{n_i-2}{n_i^2(n_i-1)}\sum_{k,k_1}^{n_i}e_{ijk}e_{ij_1k_1} - \frac{1}{n_i(n_i-1)}\sum_{k=1}^{n_i}e_{ijk}e_{ij_1k} + \right. \\
&\quad \left. \frac{2}{n_i^2(n_i-1)}\sum_{k,k_1}^{n_i}e_{ij_1k}e_{ijk_1}\right] \\
&= \frac{1}{J\sqrt{I}}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_i(n_i-1)}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ij_1k_1}.
\end{aligned}$$

Therefore, $E[Q_A]=0$. It follows that

$$\begin{aligned}
& Var(Q_A(e)) \\
&= \frac{1}{IJ^2}\sum_{i=1}^{I}\frac{1}{n_i^2(n_i-1)^2}Var\left[\sum_{j,j_1}^{J}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ij_1k_1}\right] \\
&= \frac{1}{IJ^2}\sum_{i=1}^{I}\frac{2}{n_i^2(n_i-1)^2}\sum_{k\neq k_1}^{n_i}Var\left[\sum_{j,j_1}^{J}e_{ijk}e_{ij_1k_1}\right] \\
&= \frac{1}{IJ^2}\sum_{i=1}^{I}\frac{2}{n_i^2(n_i-1)^2}\sum_{k\neq k_1}^{n_i}\sum_{j,j_1,j_2,j_3}^{J}\sigma_{i,jj_1}\sigma_{i,j_2j_3} \\
&= \frac{1}{IJ^2}\sum_{i=1}^{I}\frac{2}{n_i(n_i-1)}\sum_{j,j_1,j_2,j_3}^{J}\sigma_{i,jj_1}\sigma_{i,j_2j_3}.
\end{aligned}$$

Since $Var(Q_A(e))$ is bounded away from zero and infinity, Lyapunov's condition will be

satisfied if

$L_A(a) = \sum_{i=1}^{I} E|\frac{1}{J\sqrt{I}} \sum_{j,j_1}^{J} \frac{1}{n_i(n_i-1)} \sum_{k\neq k_1}^{n_i} e_{ijk}e_{ij_1k_1}|^4 \to 0.$

We have

$$\begin{aligned}
L_A(a) &= \frac{1}{I^2J^4} \sum_{i=1}^{I} E \left| \sum_{j,j_1}^{J} \frac{1}{n_i(n_i-1)} \sum_{k\neq k_1}^{n_i} e_{ijk}e_{ij_1k_1} \right|^4 \\
&\leq \frac{J^6}{I^2J^4} \sum_{i=1}^{I} \sum_{j,j_1}^{J} E \left| \frac{1}{n_i(n_i-1)} \sum_{k\neq k_1}^{n_i} e_{ijk}e_{ij_1k_1} \right|^4 \\
&= \frac{J^2}{I^2} \sum_{i=1}^{I} \sum_{j,j_1}^{J} \frac{1}{n_i^4(n_i-1)^4} E \left| \sum_{k\neq k_1}^{n_i} e_{ijk}e_{ij_1k_1} \right|^4 \\
&\leq \frac{J^2}{I^2} \sum_{i=1}^{I} \sum_{j,j_1}^{J} \frac{n_i^3(n_i-1)^3}{n_i^4(n_i-1)^4} \sum_{k\neq k_1}^{n_i} E|e_{ijk}e_{ij_1k_1}|^4 \\
&= \frac{J^2}{I^2} \sum_{i=1}^{I} \sum_{j,j_1}^{J} \frac{1}{n_i(n_i-1)} \sum_{k\neq k_1}^{n_i} E[e_{ijk}^4]E[e_{ij_1k_1}^4] \\
&= O(I^{-1}), \text{ if the fourth moment of } e_{ijk} \text{ exists for any } i, j, \text{ and } k.
\end{aligned}$$

where the two inequalities follow Hölder's inequality (2.5.2). This completes the proof.

Theorem 2.5.4. *For testing $H_0(B_G)$: $L\beta = \mathbf{0}$ where L is a $J \times p$ contrast matrix, $\beta = (\beta_1, ..., \beta_J)'$, and $\mathbf{0}$ is a p dimensional zero vector, let W_B be the statistic given in (2.4.3). If X_{ijk} has finite second and fourth moments, then under $H_0(B_G)$,*

$$W_B \xrightarrow{d} \chi_p^2$$

holds for all $n_i \geq 2$, i=1, ..., I.

Proof of Theorem 2.5.4: Under $H_0(B_G)$, $LE[D(B)] = \mathbf{0}$, then $LD(B) = L(D(B) - E[D(B)])$. Let $V(B) = Var[D(B)]$. V(B) is a $J \times J$ matrix, where the value of j_1th row and j_2th column is defined as

$$Cov(\widetilde{X}_{\cdot j_1 \cdot}, \widetilde{X}_{\cdot j_2 \cdot}) = \eta(B)_{j_1 j_2} = \frac{1}{I^2} \sum_{i}^{I} \frac{\sigma_{i,j_1 j_2}}{n_i}.$$

If $j_1 = j_2 = j$, it is the variance of $\widetilde{X}_{\cdot j \cdot}$, and it is denoted as

$$\eta(B)_j = \frac{1}{I^2} \sum_i^I \frac{\sigma_{i,j}^2}{n_i}.$$

The result will follow from the Continuous Mapping and Slutsky's Theorems, by showing $\sqrt{N}(D(B) - E[D(B)]) \xrightarrow{d} N_J(\mathbf{0}, NV(B))$ as $I \to \infty$, where $N = \sum_{i=1}^I n_i$. It is sufficient to show for any finite constants $\mathbf{a} = (a_1, ..., a_J)'$,

$$\sqrt{N}\mathbf{a}'(D(B) - E[D(B)]) = \sum_{j=1}^J a_j \sqrt{N}(\widetilde{X}_{\cdot j \cdot} - E(\widetilde{X}_{\cdot j \cdot})) \xrightarrow{d} N(0, N\mathbf{a}'V(B)\mathbf{a}).$$

where $N\mathbf{a}'V(B)\mathbf{a} = N\sum_{i=1}^I \sum_{j,j_1}^J a_j a_{j_1} \sigma_{i,jj_1}/n_i I^2$ converges if $(\sum_i^I \frac{1}{n_i})(\sum_i^I n_i)/I^2$ converges, which holds since n_i are finite.

$$\begin{aligned}
& \sum_{j=1}^J a_j \sqrt{N}[\widetilde{X}_{\cdot j \cdot} - E(\widetilde{X}_{\cdot j \cdot})] \\
= & \frac{\sqrt{N}}{I} \sum_i^I \sum_{j=1}^J a_j [\overline{X}_{ij \cdot} - E(\overline{X}_{ij \cdot})] \\
= & \sum_i^I \left(\frac{\sqrt{N}}{I} \sum_j^J a_j \overline{e}_{ij \cdot} \right).
\end{aligned}$$

The asymptotic normality can be shown by Lyapounov's Theorem. The Lyapounov condition will be satisfied if

$$L(B) = \sum_i^I \left(\frac{\sqrt{N}}{I} \right)^4 E \left(\sum_j^J a_j \overline{e}_{ij \cdot} \right)^4 \to 0.$$

Note that

$$\begin{aligned}
L(B) &= \sum_i^I \left(\frac{\sqrt{N}}{I} \right)^4 E \left(\sum_j^J a_j \overline{e}_{ij \cdot} \right)^4 \leq \sum_i^I \frac{N^2}{I^4} J^3 \sum_j^J E(I_j^4 \overline{e}_{ij \cdot}^4) \\
&= \sum_i^I \frac{N^2}{I^4} J^3 \sum_j^J I_j^4 \frac{1}{n_i^4} E \left(\sum_k^{n_i} e_{ijk} \right)^4 \leq \sum_i^I \frac{N^2}{I^4} J^3 \sum_j^J \frac{I_j^4}{n_i^4} B_4 \left[\sum_k^{n_i} E(e_{ijk}^2) \right]^2 \\
&= \sum_i^I \frac{N^2}{I^4} J^3 \sum_j^J \frac{I_j^4}{n_i^4} B_4 n_i^2 [E(e_{ij1}^2)]^2 = O\left(\sum_i^I \frac{N^2}{I^4 n_i^2} \right) \\
&= O(I^{-1}),
\end{aligned}$$

where the first inequality follows Hölder's inequality (2.5.2), and the last equality holds if n_i are finite. The second inequality follows from the Khintchine inequality:

Let $\{z_n\}_{n=1}^N$ be i.i.d random variables with zero mean. Let $0 < p < \infty$, then

$$\left(E|\sum_{n=1}^{N} z_n|^p \right)^{1/p} \leq B_p \left(\sum_{n=1}^{N} E|x_n|^2 \right)^{1/2}, \tag{2.5.4}$$

for some constant $B_p > 0$ depending only on p (Newman 1975). This complete the proof.

Theorem 2.5.5. *For testing $H_0(AB)$: all $\gamma_{ij} = 0$, let $F_X(AB)$ be the statistic given in (2.4.4) with $X_{ij} = X_{ij}$. If X_{ijk} has the finite fourth moment, then under $H_0(AB)$,*

$$\frac{\sqrt{I}(F_X(AB) - 1)}{V_{AB}} \xrightarrow{d} N(0,1), \quad \textit{where } V_{AB} \textit{ is defined in (2.5.5).}$$

The variance component is calculated by

$$V_{AB} = \sqrt{\tau_{AB}}/\sigma_{AB}, \tag{2.5.5}$$

where

$$\begin{aligned}
\tau_{AB} &= \frac{2}{I(J-1)^2} \sum_{i}^{I} \left[\frac{1}{n_i(n_i-1)} \sum_{j,j_1}^{J} \sigma_{i,jj_1}^2 + \frac{1}{J^2 n_i(n_i-1)} \sum_{j,j_1,j_2,j_3}^{J} \sigma_{i,jj_1}\sigma_{i,j_2j_3} - \right. \\
&\quad \left. \frac{2}{Jn_i(n_i-1)} \sum_{j,j_1,j_2}^{J} \sigma_{i,jj_1}\sigma_{i,jj_2} \right], \\
\sigma_{AB} &= \frac{1}{I(J-1)} \sum_{i=1}^{I} \sum_{j}^{J} \frac{\sigma_{i,j}^2}{n_i} - \frac{1}{IJ(J-1)} \sum_{i=1}^{I} \sum_{j,j_1}^{J} \frac{\sigma_{i,jj_1}}{n_i}.
\end{aligned}$$

Lemma 2.5.6. *Under the settings and assumptions of Theorem 2.5.5,*

$$MSE_{AB} - \sigma_{AB} \xrightarrow{p} 0 \text{ as } I \to \infty.$$

Proof:

As shown in lemma 2.5.2,

$$E\left[(X_{ijk} - \overline{X}_{ij.})(X_{ij_1k} - \overline{X}_{ij_1.})\right] = \frac{n_i - 1}{n_i}\sigma_{i,jj_1}.$$

Then

$$\begin{aligned} E(MSE_{AB}) &= \frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j}^{J}\frac{1}{n_i(n_i-1)}\sum_{k}^{n_i}E\left[(X_{ijk}-\overline{X}_{ij.})^2\right]- \\ &\quad \frac{1}{IJ(J-1)}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_i(n_i-1)}\sum_{k}^{n_i}E\left[(X_{ijk}-\overline{X}_{ij.})(X_{ij_1k}-\overline{X}_{ij_1.})\right] \\ &= \frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j}^{J}\frac{\sigma_{i,j}^2}{n_i}-\frac{1}{IJ(J-1)}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{\sigma_{i,jj_1}}{n_i} \\ &= \sigma_{AB}. \end{aligned}$$

And, we have,

$$\begin{aligned} & Var(MSE_{AB}) \\ = & \frac{1}{I^2(J-1)^2}\sum_{i=1}^{I}\frac{1}{n_i^2(n_i-1)^2}\sum_{k}^{n_i}\left\{Var\left[\sum_{j}^{J}(X_{ijk}-\overline{X}_{ij.})^2\right]+\right. \\ & \frac{1}{J^2}Var\left[\sum_{j,j_1}^{J}(X_{ijk}-\overline{X}_{ij.})(X_{ij_1k}-\overline{X}_{ij_1.})\right]- \\ & \left.\frac{2}{J}Cov\left[\sum_{j}^{J}(X_{ijk}-\overline{X}_{ij.})^2,\sum_{j,j_1}^{J}(X_{ijk}-\overline{X}_{ij.})(X_{ij_1k}-\overline{X}_{ij_1.})\right]\right\}. \end{aligned}$$

In the proof of lemma (2.5.2), we showed in formula (2.5.3) that

$$\begin{aligned} & \left|Cov[\sum_{j,j_1}^{J}(X_{ijk}-\overline{X}_{ij.})(X_{ij_1k}-\overline{X}_{ij_1.}),\sum_{j_2,j_3}^{J}(X_{ij_2k}-\overline{X}_{ij_2.})(X_{ij_3k}-\overline{X}_{ij_3.})]\right| \\ = & \left|Cov[\sum_{j,j_1}^{J}(e_{ijk}-\overline{e}_{ij.})(e_{ij_1k}-\overline{e}_{ij_1.}),\sum_{j_2,j_3}^{J}(e_{ij_2k}-\overline{e}_{ij_2.})(e_{ij_3k}-\overline{e}_{ij_3.})]\right| \\ < & \infty. \end{aligned}$$

This follows from

$$Var\left[\sum_{j}^{J}(X_{ijk}-\overline{X}_{ij.})^2\right]<\infty,$$

$$Var\left[\sum_{j,j_1}^{J}(X_{ijk}-\overline{X}_{ij.})(X_{ij_1k}-\overline{X}_{ij_1.})\right]<\infty,$$

$$Cov\left[\sum_{j}^{J}(X_{ijk}-\overline{X}_{ij.})^2,\sum_{j,j_1}^{J}(X_{ijk}-\overline{X}_{ij.})(X_{ij_1k}-\overline{X}_{ij_1.})\right]<\infty.$$

Therefore, we have $Var(MSE_{AB}) = O(I^{-1})$ and it follows that $MSE_{AB} - \sigma^2_{AB} \xrightarrow{p} 0$ as $I \to \infty$.

Lemma 2.5.7. *Under the settings and assumptions of Theorem 2.5.5 and under $H_0(AB)$, we have*

$$\sqrt{I}(MST_{AB} - P_{AB}(e)) \xrightarrow{p} 0 \text{ as } I \to \infty,$$

where $P_{AB}(e) = \frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{e_{ij.}} - \tilde{e_{i..}})^2$.

Proof:

Note that under $H_0(AB)$,

$$\begin{aligned}
& MST_{AB} \\
= & \frac{1}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\left[(\overline{e}_{ij.}-\tilde{e}_{i..})^2-\frac{2}{I}\sum_{i_1}^{I}(\overline{e}_{ij.}-\tilde{e}_{i..})(\overline{e}_{i_1j.}-\tilde{e}_{i_1..})+\frac{1}{I^2}\sum_{i_1}^{I}(\overline{e}_{i_1j.}-\tilde{e}_{i_1..})^2\right] \\
= & \frac{1}{(I-1)(J-1)}\sum_{j=1}^{J}\left[\frac{I+1}{I}\sum_{i=1}^{I}(\overline{e}_{ij.}-\tilde{e}_{i..})^2-\frac{2}{I}\sum_{i,i_1}^{I}(\overline{e}_{ij.}-\tilde{e}_{i..})(\overline{e}_{i_1j.}-\tilde{e}_{i_1..})\right] \\
= & \frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{e}_{ij.}-\tilde{e}_{i..})^2-\frac{2}{I(I-1)(J-1)}\sum_{i\neq i_1}^{I}\sum_{j=1}^{J}(\overline{e}_{ij.}-\tilde{e}_{i..})(\overline{e}_{i_1j.}-\tilde{e}_{i_1..}).
\end{aligned}$$

Thus, we have

$$E[\sqrt{I}(MST_{AB} - P_{AB}(e))] = \frac{2\sqrt{I}}{I(I-1)(J-1)}\sum_{i\neq i_1}^{I}\sum_{j=1}^{J}E\left[(\overline{e}_{ij.}-\tilde{e}_{i..})(\overline{e}_{i_1j.}-\tilde{e}_{i_1..})\right] = 0.$$

And

$$
\begin{aligned}
& E[\sqrt{I}(MST_{AB} - P_{AB}(e))]^2 \\
= \; & \frac{4I}{I^2(I-1)^2(J-1)^2} E\left[\sum_{i\neq i_1}^{I}\sum_{j=1}^{J}(\overline{e}_{ij.} - \tilde{e}_{i..})(\overline{e}_{i_1j.} - \tilde{e}_{i_1..})\right]^2 \\
= \; & \frac{4I}{I^2(I-1)^2(J-1)^2} E\left[2\sum_{(i=i_2)\neq(i_1=i_3)}^{I}\left(\sum_{j=1}^{J}(\overline{e}_{ij.} - \tilde{e}_{i..})(\overline{e}_{i_1j.} - \tilde{e}_{i_1..})\right)\right. \\
& \left.\left(\sum_{j_1=1}^{J}(\overline{e}_{i_2j_1.} - \tilde{e}_{i_2..})(\overline{e}_{i_3j_1.} - \tilde{e}_{i_3..})\right)\right] \\
= \; & \frac{8I}{I^2(I-1)^2(J-1)^2} E\left[\sum_{i\neq i_1}^{I}\sum_{j,j_1}^{J}(\overline{e}_{ij.} - \tilde{e}_{i..})(\overline{e}_{i_1j.} - \tilde{e}_{i_1..})(\overline{e}_{ij_1.} - \tilde{e}_{i..})(\overline{e}_{i_1j_1.} - \tilde{e}_{i_1..})\right] \\
= \; & \frac{8I}{I^2(I-1)^2(J-1)^2}\sum_{i\neq i_1}^{I}\sum_{j,j_1}^{J} E[(\overline{e}_{ij.} - \tilde{e}_{i..})(\overline{e}_{ij_1.} - \tilde{e}_{i..})]E[(\overline{e}_{i_1j.} - \tilde{e}_{i_1..})(\overline{e}_{i_1j_1.} - \tilde{e}_{i_1..})] \\
= \; & O(I^{-1}).
\end{aligned}
$$

Therefore under $H_0(AB)$, $\sqrt{I}(MST_{AB} - P_{AB}(e)) \xrightarrow{p} 0$ as $I \to \infty$.

Proof of Theorem 2.5.5: By Lemma 2.5.6 and Lemma 2.5.7, we need only to consider the asymptotic distribution of $Q_{AB}(e) = \sqrt{I}(P_{AB}(e) - MSE_{AB})$ under $H_0(AB)$.

With some simple algebra, we have

$$
\begin{aligned}
& Q_{AB}(e) \\
= \; & \sqrt{I}\left[\frac{1}{I(J-1)}\sum_{i}^{I}\sum_{j}^{J}(\overline{e}_{ij\cdot}-\widetilde{e}_{i\cdot\cdot})^2-\frac{1}{I(J-1)}\sum_{i}^{I}\sum_{j}^{J}\frac{1}{n_i(n_i-1)}\sum_{k}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})^2+\right. \\
& \left.\frac{1}{IJ(J-1)}\sum_{i}^{I}\sum_{j,j_1}^{J}\frac{1}{n_i(n_i-1)}\sum_{k}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})(e_{ij_1k}-\overline{e}_{ij_1\cdot})\right] \\
= \; & \frac{1}{\sqrt{I}(J-1)}\sum_{i}^{I}\left[\sum_{j}^{J}(\overline{e}_{ij\cdot}-\widetilde{e}_{i\cdot\cdot})^2-\sum_{j}^{J}\frac{1}{n_i(n_i-1)}\sum_{k}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})^2+\right. \\
& \left.\sum_{j,j_1}^{J}\frac{1}{Jn_i(n_i-1)}\sum_{k}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})(e_{ij_1k}-\overline{e}_{ij_1\cdot})\right] \\
= \; & \frac{1}{\sqrt{I}(J-1)}\sum_{i}^{I}\left[(\sum_{j}^{J}\overline{e}_{ij\cdot}^2-\frac{1}{J}\sum_{j,j_1}^{J}\overline{e}_{ij\cdot}\overline{e}_{ij_1\cdot})-\sum_{j}^{J}\frac{1}{n_i(n_i-1)}\sum_{k}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})^2+\right. \\
& \left.\sum_{j,j_1}^{J}\frac{1}{Jn_i(n_i-1)}\sum_{k}^{n_i}(e_{ijk}-\overline{e}_{ij\cdot})(e_{ij_1k}-\overline{e}_{ij_1\cdot})\right] \\
= \; & \frac{1}{\sqrt{I}(J-1)}\sum_{i}^{I}\left[(\frac{1}{n_i^2}\sum_{j}^{J}\sum_{k,k_1}^{n_i}e_{ijk}e_{ijk_1}-\frac{1}{Jn_i^2}\sum_{j,j_1}^{J}\sum_{k,k_1}^{n_i}e_{ijk}e_{ij_1k_1})-\right. \\
& (\frac{1}{n_i(n_i-1)}\sum_{j}^{J}\sum_{k}^{n_i}e_{ijk}^2-\frac{1}{n_i^2(n_i-1)}\sum_{j}^{J}\sum_{k,k_1}^{n_i}e_{ijk}e_{ijk_1})+ \\
& \left.(\frac{1}{Jn_i(n_i-1)}\sum_{j,j_1}^{J}\sum_{k}^{n_i}e_{ijk}e_{ij_1k}-\frac{1}{Jn_i^2(n_i-1)}\sum_{j,j_1}^{J}\sum_{k,k_1}^{n_i}e_{ijk}e_{ij_1k_1})\right] \\
= \; & \frac{1}{\sqrt{I}(J-1)}\sum_{i}^{I}\left[\frac{1}{n_i(n_i-1)}\sum_{j}^{J}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ijk_1}-\frac{1}{Jn_i(n_i-1)}\sum_{j,j_1}^{J}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ij_1k_1}\right].
\end{aligned}
$$

Therefore, $E[Q_{AB}(e)] = 0$. It follows that

$$
\begin{aligned}
& Var(Q_{AB}(e)) \\
= \; & \frac{1}{I(J-1)^2} \sum_{i}^{I} Var \left[\frac{1}{n_i(n_i-1)} \sum_{j}^{J} \sum_{k \neq k_1}^{n_i} e_{ijk} e_{ijk_1} - \frac{1}{Jn_i(n_i-1)} \sum_{j,j_1}^{J} \sum_{k \neq k_1}^{n_i} e_{ijk} e_{ij_1k_1} \right] \\
= \; & \frac{2}{I(J-1)^2} \sum_{i}^{I} \sum_{k \neq k_1}^{n_i} \left[Var(\frac{1}{n_i(n_i-1)} \sum_{j}^{J} e_{ijk} e_{ijk_1}) + Var(\frac{1}{Jn_i(n_i-1)} \sum_{j,j_1}^{J} e_{ijk} e_{ij_1k_1}) - \right. \\
& \left. 2Cov(\frac{1}{n_i(n_i-1)} \sum_{j}^{J} e_{ijk} e_{ijk_1}, \frac{1}{Jn_i(n_i-1)} \sum_{j,j_1}^{J} e_{ijk} e_{ij_1k_1}) \right] \\
= \; & \frac{2}{I(J-1)^2} \sum_{i}^{I} \sum_{k \neq k_1}^{n_i} \left[\frac{1}{n_i^2(n_i-1)^2} \sum_{j,j_1}^{J} E(e_{ijk} e_{ij_1k}) E(e_{ijk_1} e_{ij_1k_1}) + \right. \\
& \left. \frac{1}{J^2 n_i^2 (n_i-1)^2} \sum_{j,j_1,j_2,j_3}^{J} E(e_{ijk} e_{ij_2k}) E(e_{ij_1k_1} e_{ij_3k_1}) - \frac{2}{Jn_i^2(n_i-1)^2} \sum_{j,j_1,j_2}^{J} E(e_{ijk} e_{ijk_1} e_{ij_1k} e_{ij_2k_1}) \right] \\
= \; & \frac{2}{I(J-1)^2} \sum_{i}^{I} \left[\frac{1}{n_i(n_i-1)} \sum_{j,j_1}^{J} \sigma_{i,jj_1}^2 + \frac{1}{J^2 n_i(n_i-1)} \sum_{j,j_1,j_2,j_3}^{J} \sigma_{i,jj_1} \sigma_{i,j_2j_3} - \right. \\
& \left. \frac{2}{Jn_i(n_i-1)} \sum_{j,j_1,j_2}^{J} \sigma_{i,jj_1} \sigma_{i,jj_2} \right].
\end{aligned}
$$

Since $Var(Q_{AB}(e))$ is bounded away from 0 and ∞, Lyapounov's condition will be satisfied if

$$
\begin{aligned}
L_{AB}(a) \; &= \; \sum_{i=1}^{I} E \left| \frac{1}{\sqrt{I}(J-1)} [(\frac{1}{n_i(n_i-1)} \sum_{j}^{J} \sum_{k \neq k_1}^{n_i} e_{ijk} e_{ijk_1} - \frac{1}{Jn_i(n_i-1)} \sum_{j,j_1}^{J} \sum_{k \neq k_1}^{n_i} e_{ijk} e_{ij_1k_1})] \right|^4 \\
&\rightarrow \; 0.
\end{aligned}
$$

We then have that

$$
\begin{aligned}
L_{AB}(a) &= \frac{1}{I^2(J-1)^4}\sum_{i=1}^{I}\frac{1}{n_i^4(n_i-1)^4}E\left|(\sum_{j}^{J}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ijk_1}-\frac{1}{J}\sum_{j,j_1}^{J}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ij_1k_1})\right|^4 \\
&\leq \frac{8}{I^2(J-1)^4}\sum_{i=1}^{I}\frac{1}{n_i^4(n_i-1)^4}\left[E(\sum_{j}^{J}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ijk_1})^4+E(\frac{1}{J}\sum_{j,j_1}^{J}\sum_{k\neq k_1}^{n_i}e_{ijk}e_{ij_1k_1}))^4\right] \\
&\leq \frac{8}{I^2(J-1)^4}\sum_{i=1}^{I}\frac{n_i^3(n_i-1)^3}{n_i^4(n_i-1)^4}\sum_{k\neq k_1}^{n_i}\left[E(\sum_{j}^{J}e_{ijk}e_{ijk_1})^4+\frac{1}{J^4}E(\sum_{j,j_1}^{J}e_{ijk}e_{ij_1k_1}))^4\right] \\
&\leq \frac{8}{I^2(J-1)^4}\sum_{i=1}^{I}\frac{n_i^3(n_i-1)^3}{n_i^4(n_i-1)^4}\sum_{k\neq k_1}^{n_i}\left[J^3\sum_{j}^{J}E(e_{ijk}e_{ijk_1})^4+\frac{J^6}{J^4}\sum_{j,j_1}^{J}E(e_{ijk}e_{ij_1k_1})^4\right] \\
&= \frac{8}{I^2(J-1)^4}\sum_{i=1}^{I}\frac{n_i^3(n_i-1)^3}{n_i^4(n_i-1)^4}\sum_{k\neq k_1}^{n_i}\left[J^3\sum_{j}^{J}E(e_{ijk}^4)E(e_{ijk_1}^4)+J^2\sum_{j,j_1}^{J}E(e_{ijk}^4)E(e_{ij_1k_1}^4)\right] \\
&= O(I^{-1})\ if\ the\ fourth\ moment\ of\ e_{ijk}\ exist\ for\ any\ i,\ j,\ and\ k,
\end{aligned}
$$

where the two inequalities follow Hölder's inequality (2.5.2). This completes the proof.

2.6 Simulation results

In order to evaluate the proposed non-parametric test statistics (NPT), we compare NPT with linear mixed estimation model (LME) and generalized estimating equations (GEE) by simulation studies. First, we performed simulation to calculate the type I error rates for random numbers generated from various distributions and covariance structures. Secondly, we generate bootstrap re-sampling data from real aCGH profiles. We then introduced within-subject correlation to the data, and conducted power analysis to compare NPT, LME, and GEE. All the data in this section were generated from the model specified in (2.3.1)

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}.$$

We used R programming to conduct all calculations and simulations. For calculations based on LME and GEE methods, R packages *nlme* and *geepack* were used.

2.6.1 Simulated data

In this subsection, simulations were used to estimate type I error rates for the proposed method (NPT). The aCGH data were often modeled with log-normal or Poisson distribution (Sidorov et al. (2002); Zhao et al. (2004)). Nonetheless, such models have been shown to be inappropriate because they skew data histograms or probability plots (Konishi 2004). Therefore, we used normal, exponential, Poisson, or Cauchy distributions to generate random samples. Proper within-subject correlation structures were introduced into the data as well. All simulations were conducted using 1000 iterations.

As the copy number of normal chromosomes is 2, we used a mean of 2 for normal, exponential,and Poisson distributions. The normal distribution had a standard deviation equal to 1. The Cauchy distribution had a location parameter 0, and a scale parameter 1.

The within-subject correlation (over time points) were modeled either with an AR(1) or an unstructured correlation structure. For AR(1) correlation, the covariance vector X was conditioned by $cov(X_{ijk}, X_{ij_1k}) = .5^{|j-j_1|}$. The unstructured correlation structures were obtained by generating a symmetric matrix that has random numbers uniformly distributed between 0 and 1. The methods to introduce the correlation structures are described below.

First, we examined the proposed test statistic for $H_0(A)$ of no probe copy number variation. The probe copy numbers were randomly generated under null hypothesis of equal numbers for distinct subjects at the same time point. For convenience, we used the same mean for all probes at all time points. For the balanced design, the random numbers were put in a matrix X of J rows and $I \times n$ columns. J is the number of time points, I is the number of probes, and n is the number of replications. For unbalanced design, the number of columns of X is the sum of the number of replications for all individual probes.

An AR(1) or unstructured correlation structure $J \times J$ matrix L was then generated as described above. The Cholesky decomposition of L calculates the lower half triangle matrix

h. Thus, we have the simulated data matrix $Y = h \cdot X$ that had the desired correlation structure. The matrix Y had equal means across columns. Nonetheless, at different time points (between rows), the copy numbers from the same probe could vary. It was then used for analysis by our statistic defined in 2.4.1. There is an exception in the data generating process for Poisson distribution. We intended to use Poisson distribution to generate integer copy number data. Therefore, we first calculated the means matrix M for Y by multiplying the half matrix h with a data matrix consisting of only copy number 2's. Then the random data matrix Y was generated by using Poisson distribution with mean M.

In the test of the probe effect, we considered both balanced and unbalanced designs. Each sample have 5 replicates in the balanced design. The type I error rates with an AR(1) correlation were shown in Table 2.1. Either 2 or 8 time points were simulated for each dataset. The number of time points does not significantly affect the error rates. They had similar error rates in all conditions. We increased the number of probes from 5 to 1000. At the level $\alpha = 0.05$, the error rate converges to 0.05 as the number of probes increases, and they were close to 0.05 for normal, exponential and Poisson distributions when the number of probes was 40 or above. The error rate for Cauchy distribution did not converge to 0.05 since Cauchy distribution does not have a finite mean.

For the unbalanced design, we created data by assigning four fifths of probes with 4 replications, and the remaining probes with 6 replications. The results of AR(1) and unstructured correlation are shown in Table 2.2 and Table 2.3 separately. The conclusions were similar to those of balanced designs (Table 2.3).

Secondly, we conducted hypothesis test for the time effect. Similar to that of the probe effect, we first generate random copy numbers matrix X with equal means for all tested distributions except for Poisson distribution. In order to maintain equal means across time points (rows), we cannot use the Cholesky decomposition to introduce correlation structure. Instead, we used a iterative algorithm. Suppose for probe i, the correlation between the jth

no.time	no.snp	normal	exponential	Poisson	Cauchy
2	5	0.107	0.117	0.118	0.190
	10	0.091	0.076	0.096	0.173
	20	0.085	0.073	0.066	0.147
	30	0.074	0.063	0.059	0.135
	40	0.063	0.060	0.048	0.132
	50	0.050	0.055	0.053	0.128
	100	0.053	0.051	0.053	0.107
	200	0.051	0.046	0.060	0.123
	500	0.054	0.053	0.056	0.111
	1000	0.042	0.052	0.049	0.104
8	5	0.103	0.098	0.105	0.206
	10	0.094	0.072	0.075	0.164
	20	0.078	0.071	0.059	0.156
	30	0.065	0.064	0.071	0.141
	40	0.059	0.042	0.063	0.128
	50	0.054	0.064	0.061	0.128
	100	0.049	0.059	0.055	0.130
	200	0.051	0.059	0.057	0.111
	500	0.049	0.059	0.035	0.103
	1000	0.051	0.051	0.063	0.085

Table 2.1: Estimated type I error estimate of the test of no probe effect at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. There are 5 replications in the design.

and (j+1)th time points is ρ. Given a copy number X_j for the jth time point, the random copy number of the (j+1)th time point can be generated by

$$X_{j+1} = \rho X_j + b,$$

where b is a random number with mean $2(1-\rho)$. Thus, the mean of X_{j+1} is 2. For Poisson distribution, similar to the process for the probe effect, we first generated the means with desired correlation structure, and then used the means to generate random integer copy numbers.

The type I error rates at alpha level 0.05 with an AR(1) correlation were shown in Table 2.4. Two or eight time points were simulated for each experiment. For each dataset,

no.snp	normal	exponential	Poisson	Cauchy
10	0.102	0.111	0.101	0.163
20	0.083	0.074	0.075	0.149
30	0.071	0.067	0.072	0.121
40	0.062	0.066	0.074	0.138
50	0.072	0.059	0.057	0.134
100	0.051	0.068	0.067	0.111
200	0.041	0.038	0.043	0.114
500	0.047	0.056	0.053	0.112
1000	0.052	0.053	0.048	0.087

Table 2.2: Estimated type I error of the test of no probe effect at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. The number of time points is 2. For each experiment, Four fifths of probes have 4 replications, and the remaining one fifth of probes have 6 replications.

no.snp	normal	exponential	Poisson	Cauchy
10	0.089	0.091	0.100	0.156
20	0.077	0.072	0.075	0.133
30	0.060	0.072	0.054	0.117
40	0.071	0.069	0.048	0.120
50	0.070	0.071	0.043	0.116
100	0.054	0.074	0.051	0.121
200	0.055	0.057	0.048	0.121
500	0.050	0.055	0.057	0.108
1000	0.046	0.063	0.053	0.100

Table 2.3: Estimated type I error of the test of no probe effect at 0.05 level. The data from the same probe follow unstructured correlation. The number of time points is 5. For each experiment, Four fifths of probes have 4 replications, and the remaining one fifth of probes have 6 replications.

four fifths of probes were assigned 4 replications, and the remaining one fifth of probes were assigned 6 replications. As the number of time points increases, it needs more probes to reach the expected error rate. Normal, exponential, and Poisson distributions showed similar convergence rate. But as expected, the error rate for Cauchy distribution did not converge to 0.05.

Table 2.5 showed the type I error rates for the unstructured correlation. The conclusions

no.time	no.snp	normal	exponential	Poisson	Cauchy
2	10	0.060	0.061	0.051	0.030
	20	0.064	0.052	0.047	0.029
	30	0.067	0.054	0.063	0.025
	40	0.053	0.050	0.052	0.024
	50	0.053	0.054	0.058	0.011
	100	0.043	0.055	0.051	0.022
	200	0.048	0.047	0.048	0.017
	500	0.044	0.045	0.053	0.019
	1000	0.058	0.037	0.052	0.016
8	5	0.373	0.404	0.349	0.319
	10	0.192	0.207	0.176	0.106
	20	0.113	0.125	0.106	0.052
	30	0.092	0.090	0.077	0.033
	40	0.075	0.075	0.088	0.024
	50	0.065	0.082	0.074	0.027
	100	0.048	0.059	0.049	0.017
	200	0.052	0.044	0.055	0.011
	500	0.052	0.053	0.042	0.015
	1000	0.055	0.052	0.039	0.011

Table 2.4: Estimated type I error estimate of the test of no time effect at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. For each probe, the number of replications is either 4 or 6.

were similar to those of the balanced designs.

Thirdly, simulation was conducted to test the interaction of probe and time. The data generating process was similar to that for test of time effect. Under null hypothesis, the copy number for all probes at all time points are equal. We gave 8 time points in the experiment. An unbalanced design was used such that four fifths of probes have 4 replications, and one fifth of probes have 6 replications. The type I error rate at alpha level 0.05 were reported in Table 2.6 and Table 2.7 for AR(1) and unstructured correlations, respectively. Normal, exponential, and Poisson distributions showed similar convergence rate. Like other tests, the error rate for Cauchy distribution did not converge.

no.snp	normal	exponential	Poisson	Cauchy
10	0.114	0.121	0.103	0.052
20	0.084	0.074	0.086	0.027
30	0.070	0.093	0.058	0.023
40	0.065	0.069	0.077	0.026
50	0.065	0.074	0.060	0.019
100	0.038	0.054	0.047	0.016
200	0.052	0.051	0.048	0.015
500	0.038	0.039	0.041	0.018
1000	0.049	0.038	0.045	0.017

Table 2.5: Estimated type I error of the test of no time effect at 0.05 level. The data from the same probe follow unstructured correlation. For each simulation, there are 8 time points. For each probe, the number of replications is either 4 or 6.

no.snp	normal	exponential	Poisson	Cauchy
10	0.086	0.084	0.101	0.175
20	0.071	0.079	0.066	0.153
30	0.073	0.076	0.074	0.149
40	0.056	0.073	0.066	0.151
50	0.045	0.051	0.054	0.142
100	0.047	0.062	0.052	0.139
200	0.070	0.054	0.056	0.105
500	0.035	0.044	0.046	0.097
1000	0.062	0.047	0.055	0.095

Table 2.6: Estimated type I error rates of the test of no interaction of probe and time effects at 0.05 level. The data from the same probe follow AR(1) with correlation =0.5. For each probe, the number of replications is either 4 or 6.

2.6.2 Bootstrap data

In this subsection, we used power analysis to compare the proposed method (NPT) with linear mixed-effects model (LME) and generalized estimating equations (GEE). To simulate data as closely as possible to the real aCGH data, we used bootstrap to generate re-sampled data based on an aCGH application.

It has been reported that amplification of chromosome 7q is associated with glioma

no.snp	normal	exponential	Poisson	Cauchy
10	0.080	0.087	0.086	0.175
20	0.076	0.074	0.057	0.155
30	0.050	0.070	0.059	0.153
40	0.048	0.057	0.055	0.145
50	0.049	0.063	0.051	0.136
100	0.060	0.052	0.043	0.117
200	0.050	0.065	0.053	0.140
500	0.050	0.071	0.047	0.110
1000	0.061	0.062	0.040	0.099

Table 2.7: Estimated type I error rates of the test of no interaction of probe and time effects at 0.05 level. The data from the same probe follow unstructured correlation with correlation =0.5. For each probe, the number of replications is either 4 or 6.

tumor (Maher (2006)). We acquired the copy numbers of 3,000 SNPs in chromosome 7q from Affymetrix 100K SNP arrays for both a healthy person and a glioma patient. From Figure 2.4, we see that the glioma sample has 7q amplification. Its mean copy number is 4.4. The normal sample has a mean of 2.05. In the simulation design, 100 SNPs were repeatedly measured at 5 time points with either a balanced or unbalanced design. In each dataset, the majority of data came from normal 7q sample (under H_0), and they were contaminated with a small proportion of glioma data (under H_a).

Figure 2.5 showed the power curves of testing SNP effects in the balanced design with an AR(1) correlation structure. In each experiment, there were 5 time points, and 5 replications. The contamination percentage of glioma SNPs varied from 0 to 2%. The re-sampling data formed a data matrix X with 5 rows and 500 (5×100) columns. Each row represented a time point, and each column represented a SNP. The 5 replications of each SNP were in adjacent columns. The AR(1) correlation structure was introduced with the Cholesky decomposition as described in the subsection 2.6.1. The proposed method (NPT) had the fastest convergence rate to 1. The power was 100% when there were at least 0.9% contamination. At 0.9% contamination, the power of LME was 54.5%, and that of GEE was only 25.3%. GEE had the worst power among the three methods. It had a power of

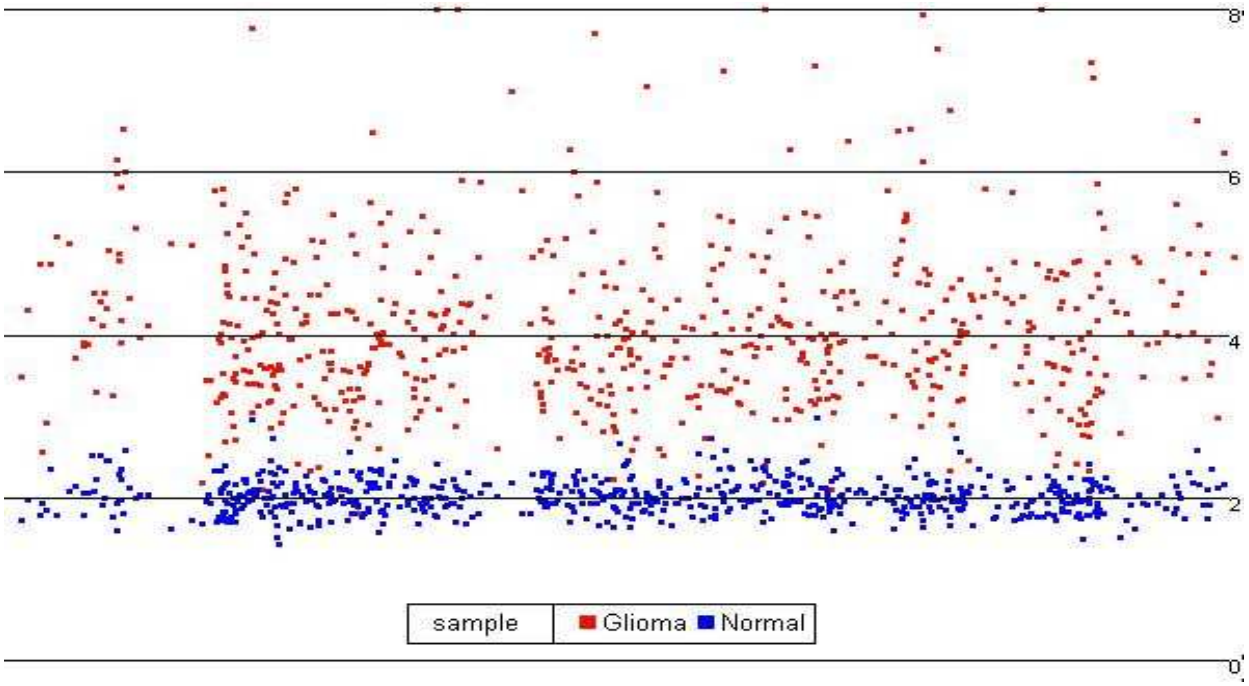


Figure 2.4: *The plots of DNA copy numbers in chromosome 7q of normal and glioma samples. Red plots denote the copy numbers of glioma SNPs, and blue plots denote the copy numbers of normal SNPs. The x axis showed the genomic positions of each SNP on chromosome 7q.*

96.9% for 2% contamination whereas the other two methods had 100% power.

We next considered unbalanced design and unstructured correlation for the bootstrap-resampled data. Four fifths of SNPs were assigned 4 replications, whereas the remaining one fifth of SNPs were assigned 6 replications. An unstructured correlation matrix were introduced with the Cholesky decomposition such that the correlation between distinct time points for the same SNP was random. Still, we considered 5 time points. The contamination percentage of glioma SNPs in each dataset was in the range of 0 to 2%. The conclusion was similar to that of the balanced design (Figure 2.6). The proposed method (NPT) outperformed the other two methods. It reached 100% power when there was at least 1.3% of contamination, whereas LME and GEE had 75% and 51% of powers, respectively. With 2% of contamination, LME had 97.3% of power, and GEE had 88.1% of power.

For the test of the time effect, bootstrap-resampled data were generated with 5 time points. The design was unbalanced and the correlation structure was unstructured. The

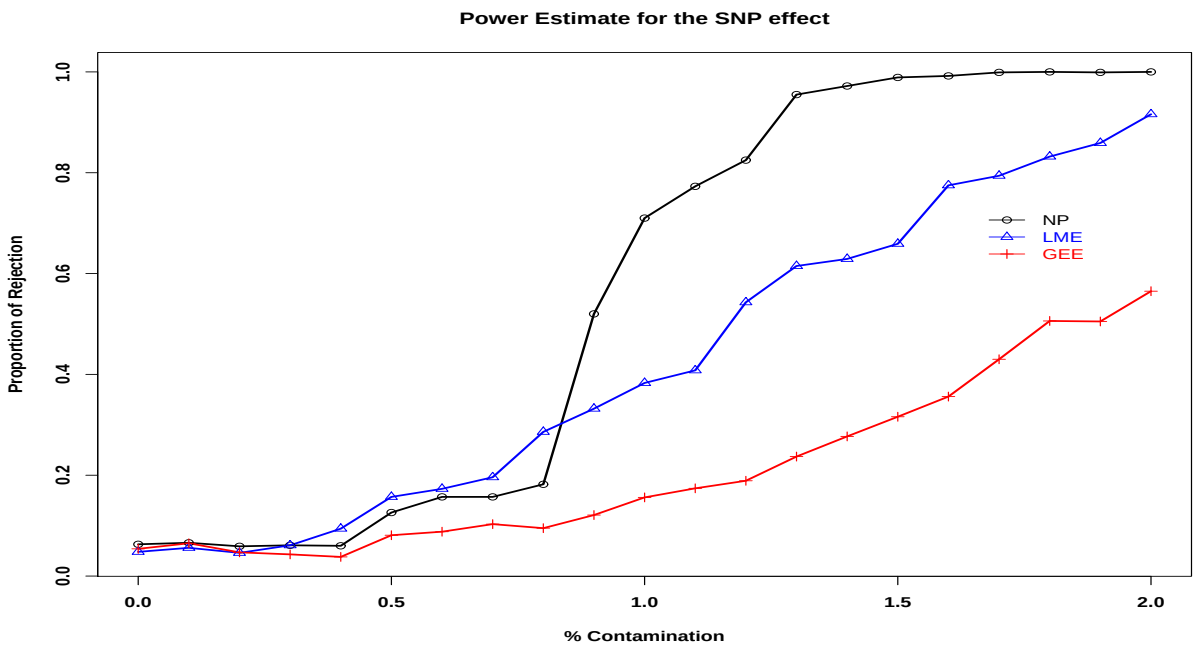


Figure 2.5: *The power curves of balanced design with an AR(1) correlation.*

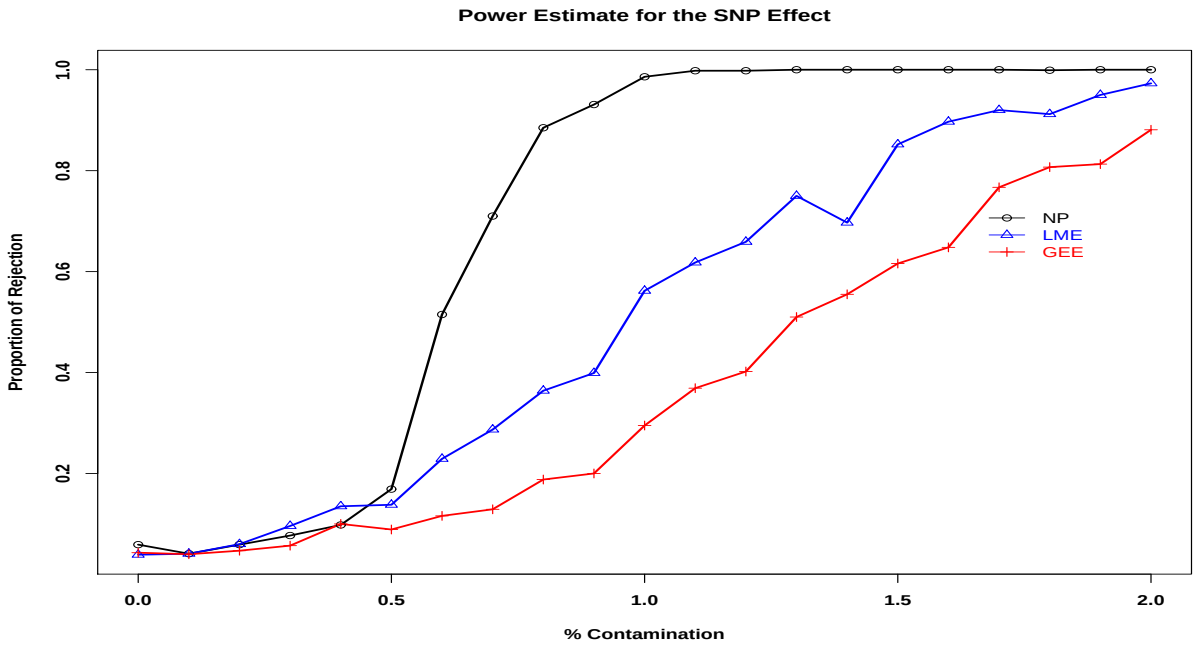


Figure 2.6: *The power curves of unbalanced design with unstructured correlation.*

correlation structure was incorporated via the interactive algorithm as described in the previous subsection. Figure 2.7 showed the power curves for the three methods. With 2%

of amplified copy number contamination, NPT had a power of 96.7%, LME of 71.8%, and GEE of 62.2%.

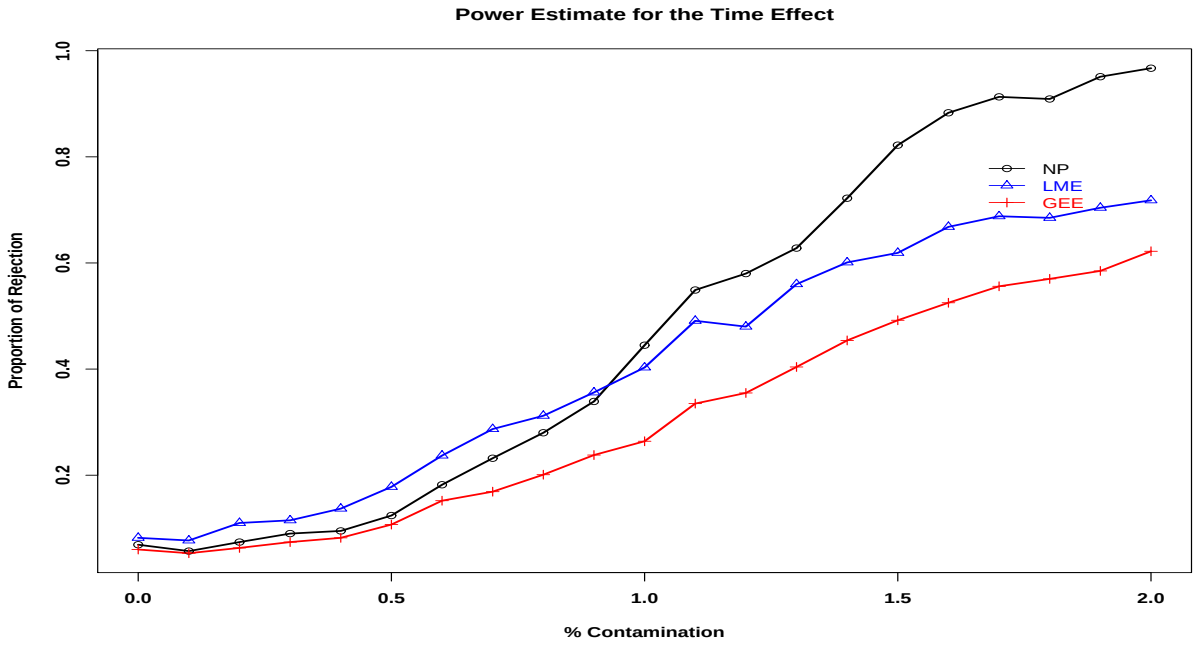


Figure 2.7: *The power curves of unbalanced design with unstructured correlation*

For the test of the interaction of SNP and time, we generated unbalanced unstructured correlated data by bootstrap-resampling for 5 time points. Similarly, we used the interactive algorithm to generate the unstructured correlation as discussed earlier. Figure 2.8 showed the power curves for the three methods. The contamination of amplified SNPs were changed from 0% to 4%. The three methods were not discriminable when the contamination is less than 0.9%. Nonetheless, NPT had a higher power for more than 0.9% of contamination. GEE performed better than LME in the interaction test, but not in the SNP and the time effect tests. With 4% of contamination, NPT had a power of 96.8%, LME of 63.6%, and GEE of 89.0%.

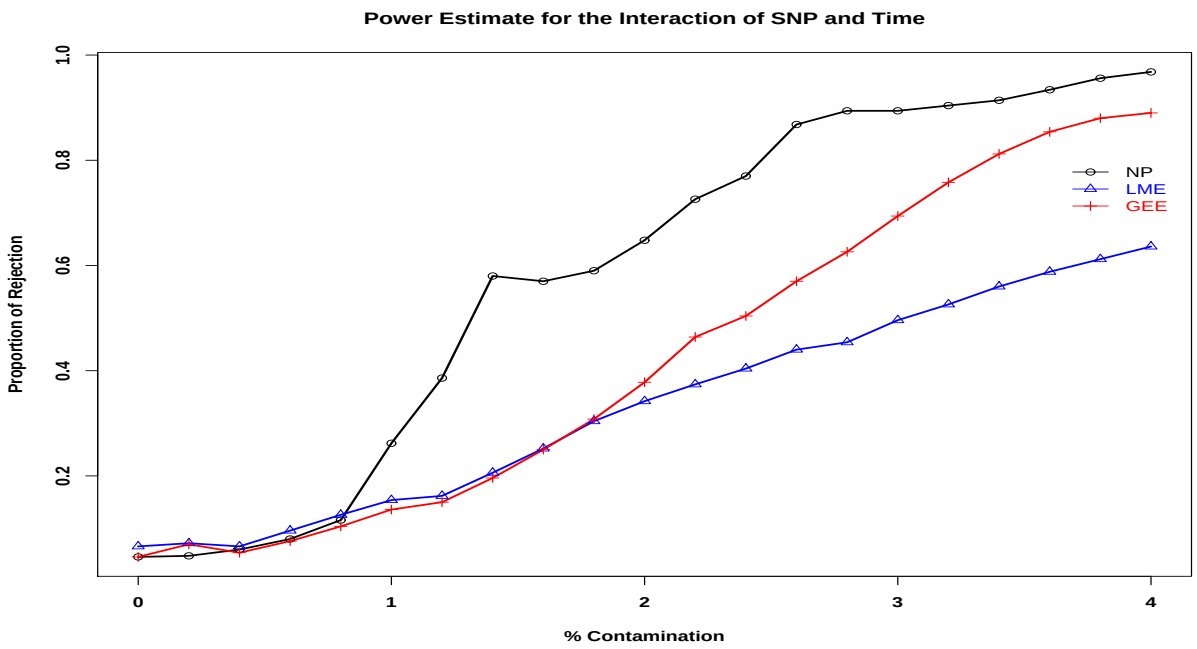


Figure 2.8: *The power curves of unbalanced design with unstructured correlation.*

2.7 A longitudinal study

Wilms' tumor typically occurs in children's kidney. Although the percentage of patients who survive at least five years is above 90%, 15% of patients will suffer from tumor relapse (Kalapurakal et al. (2004)). A lot of recent work regarding Wilms' tumor has been aimed at discovering the genetic biomarkers for diagnosis, prognosis, and treatment management (Eggert et al. (2001); Takahashi et al. (2002); Williams et al. (2004); Dome et al. (2005)). Genetic aberrations such as loss of heterozygosity and chromosome copy number changes have been found to be associated with the tumor relapse (Grundy et al. (2005); Yuan et al. (2005); Natrajan et al. (2006)). However, few longitudinal studies have been conducted to identify biomarkers that are responsible to tumor progression and recurrence.

Natrajan et al. (2007) carried out aCGH experiments for 10 Wilms' patients with relapse. The aCGH samples were conducted at both diagnosis and relapse for each patient. They used Breakthrough Breast Cancer Human CGH 4.6K 1.1.2 arrays that consist of 4179

Bacterial Artificial Chromosome (BAC) clones. The BAC clones serve as probes for measuring the genomic DNA copy number. In their report, 29 chromosome regions were identified to have copy number alterations responsible to Wilms' tumor relapse. However, their conclusions were based on pairwise comparison of diagnosis and relapse for each patient. It is not statistically justified to conduct hypothesis test without replications of subject. The reproducibility of such analysis is expected to be low, and the claimed biomarkers are possibly not useful for predicting the potential relapse of new Wilms' patients. In fact, only 6 of the 29 regions were found in 2 of the 10 patients according the their paper. Motivated by the need to redo the analysis with rigorous statistical method, we acquired the raw aCGH data and conducted analysis with the following steps.

We first performed quality control and normalization for the raw data. As female and male people have different number of sex chromosomes, to avoid them confounding with the analysis, we removed X and Y chromosomes from the data. The raw data were adjusted to baseline by subtracting the median background signal. In the experiment, each probe was labeled with two fluorescent dyes, Cy5 and Cy3. The fluorescent intensity ratio of Cy5/Cy3 were used as input data. The Cy5/Cy3 ratio were subject to quantile normalization across all samples (Bolstad et al. (2003)). The processed data had a median copy number of 2 and a standard deviation of 0.04 for each sample. They were used for subsequent analysis.

As discussed in section 2.1, a first goal of a CGH study is usually to detect the gain or loss of a chromosome arm because it is often the unit of genomic mutation and translocation activity. For instances, Hing et al. (2001) and Lu et al. (2002) found that the gain of chromosome 1q is associated with relapse of Wilms' tumor. We applied our proposed methods to each chromosome arm for hypothesis tests of probe, time, and probe$\times$ time effects. There were no probes for 5 of the 44 arms. They were 13p, 14p, 15p, 21p, and 22p. For the other arms, the minimum number of probes was 84, and the maximum number was 699. Table 2.8 lists the chromosome arms that were statistically significant for the tests. Totally 16 arms showed significant probe$\times$ time interaction. That implied the copy numbers

of some of the probes in these arms varied between diagnosis and relapse. Of the remaining chromosome arms, no time effects were detected, and two arms showed significant probe effects.

	p arm			q arm		
chromosome	probe	time	probe×time	probe	time	probe×time
1						
2	2.552E−03		2.720E−03			
3	6.697E−03		2.472E−04	0.014		0.010
4						
5	5.070E−09		1.216E−11	2.382E−04		9.936E−05
6						0.034
7	2.426E−08		0.023			0.031
8	0.041					
9	0.036		0.036			
10						
11				5.054E−08		4.350E−07
12			0.016			
13						
14						
15						0.038
16				5.538E−05		6.836E−07
17						0.015
18	2.554E−14		1.776E−15			
19						
20						
21				1.593E−03		
22						0.013

Table 2.8: Summary of significant P values (< 0.05) calculated by NPT methods for each chromosome arm.

We were most interested in the chromosome regions in which all probes had copy number gain or loss simultaneously and had no significant effect. From Table 2.8, 26 chromosome arms were not detected for any effect. We calculated the mean value for each of the arm. Unfortunately, none of these mean values was abnormally higher or lower than 2. Further analysis can be conducted by comparing to normal reference samples with the non-parametric methods proposed by Wang and Akritas (2004). The desired biomarkers for predicting relapse should show a consistent pattern between diagnosis and relapse. If

a genetic event only occurs in either of the two measures, its association with tumor recurrence is hard to be established. For the purpose of identifying prognostic biomarkers, we were not interested in the chromosome arms with probe×time interaction. Nevertheless, the interaction may indicate important genetic regulation mechanisms, and be worth further biological studies.

We explored the chromosome 8p and 21q that showed only probe effect. Significant probe effect suggests some regions in the two arms have gain or loss of DNA copies. By calculating the mean value of each probe with the measures from both diagnosis and relapse, we found four regions with abnormal copy numbers. The results were summarized in Table 2.9. Chromosome region 8p21.3 was found to have a DNA deletion. Two genes are encoded in this region. INTS10 is a subunit of RNA polymerase. Reduced expression level of RNA polymerase could lead to abnormal expression of many other genes. Thus, it is a potential oncogenesis gene. LPL is responsible for lipoprotein uptake, and it was reported to be associated with prostate cancer (Narita et al. (2004)). Chromosome 21q11.1 and 21q11.3 loss may affect the expression of genes CR614803, NCAM2, and CYYR1. However, the gene functions and their relevancy with cancer is not clear currently. The loss of 21q22.3 were associated with functions of 3 genes, NX1, NX2, and TMPRSS2. NX1 is responsible for anti-viral reaction; NX2 is a subunit of GTPase; TMPRSS2 belongs to the serine protease family. Both GTPase and serine protease are involved in a number of fundamental gene regulation pathways. The four selected regions overlapped with 2 copy number alterations reported by Natrajan et al. (2007). Thus, out of their 29 selected regions, we were only able to verify 2 potential biomarkers.

Longitudinal aCGH studies can provide unique insights into the genetic abnormalities involved in disease development and progression. However, there are a lot of challenges faced by statistical analysis. Researchers often use over-simplified analysis methods that are not able to provide sufficient statistical power and justification. We provided a robust tool based on non-parametric statistics that has potentially broad applications in this area.

Genomic region	Gene	Function
8p21.3	INTS10	RNA transcription
	LPL	lipoprotein
21q21.1	CR614803	NA
	NCAM2	NA
21q21.3	CYYR1	NA
21q22.3	NX1	anti-viral response
	NX2	GTPase
	TMPRSS2	Serine protease

Table 2.9: Summary of the copy number alterations detected for both primary and relapse tumors.

Chapter 3

Statistical tests for time course microarray data

Time course microarray experiments have been widely used to explore dynamic changes in gene expression in varying biological conditions. In a longitudinal gene expression study, each subject is repeatedly measured over time. Statistical models need to take within-subject correlation into account. In this chapter, we provide robust test procedures to compare groups of genes under multiple treatments or experimental conditions using expression data.

3.1 Introduction

Recent advances in high-throughput screening technology and in high-dimensional data analysis have made it possible for scientists to study more complex problems, such as measuring dynamic response of organisms at the molecule level. For examples, time-course microarray experiments have been conducted to investigate gene expression in the cell cycle, in the *Drosophila* immune response, in the mouse cardiac development, in the human osteoblast differentiation, in the inflammatory response of human blood leukocyte, and in the aging of human kidney cortex tissue (Shedden and Cooper (2002); Gregorio et al. (2001); Qi et al. (2003); Rodwell1 et al. (2004); Calvano et al. (2005)). Many time-course microarray experiments are designed to repeatedly measure the gene expression from the same object

over time. The following two examples highlight the structure of the data.

Example 1. A recent study was carried out on human (Calvano 2005) to study gene expression over time during acute inflammatory and immune response. Gene expression in whole blood leukocytes was determined by microarray immediately before and at 2, 4, 6, 9 and 24 h after the intravenous administration of bacterial endotoxin to four healthy human subjects. Four additional subjects were studied under identical conditions but without endotoxin administration. The blood samples were taken from the same patient at different time points were therefore correlated. Changes in blood leukocyte gene expression patterns were analyzed. The study provide insight into the generic regulation of global leukocyte activities.

Example 2. The Hessian fly (Mayetiola destructor) is one of the most destructive pests of wheat in the U.S., Western Asia, and Northern Europe. Resistance (R) genes in wheat have been the most effective means in controlling Hessian fly damages. The challenge for using resistance genes is that the effectiveness of a R gene is short-lived, lasting from six to eight years after its initial deployment Hatchett et al. (1987), Ratcliffe et al. (2000). Consequently, new R genes need to be continuously identified and incorporated into wheat cultivars for continued success. In addition, experiments at Kansas State University found that rice is a nonhost plant for Hessian fly since 100% Hessian fly larvae died in rice during the whole larval stage (Chen et al. 2008). To identify genes and pathways that were affected in resistant wheat, suspectable wheat and rice leading to host and nonhost resistance, whole genome arrays of wheat and rice were used at different time points (half-day, 1-day, 3-day and 5-day) after Hessian fly attacks. The genes affected at an early time are likely involved in regulations and signal transduction whereas the genes affected at a later time are likely involved in direct chemical defense. Large amount of array data from this experiment involving interactions of Hessian fly with wheat and rice over time remain to be analyzed effectively to identify critical genes for genetic engineering.

In the examples above, the temporal component is an inherent part of the study and multiple treatments are involved for discovery of important genes or transcriptional activity over developmental stages. These poses novel challenges for statistical analysis since effective methods need to take into account both curse of dimensionality and within-subject correlations. A few statistical attempts have been made to analyze longitudinal microarray data. Most of data analysis for longitudinal microarray is based on statistical test for individual genes and then is adjusted for multiple tests by false discovery rate (FDR). Linear mixed-effects model (LME) and generalized estimating equation (GEE) are commonly used for longitudinal data analysis (Liang and Zeger (1986); Diggle et al. (2002)). While they have been proved to be useful tools for repeated measures with large sample size, the adequacy of model fitting is of concern for high dimensional data analysis (Fan and Zhang (2000)). Park et al. (2003) used a two-stage ANOVA model to calculate the P values for each gene. At the first stage, the time effect is tested; the residuals of the time effect are then used for permutation test. They require the study to be balanced and there is no strong within-subject correlation. Guo et al. (2003) proposed a modified Wald statistic to test the differential expression of each gene over time. The Wald statistic converges to a χ^2 distribution under null hypothesis when the number of subjects is sufficiently large. Each gene was assigned a gene-specific score that was calculated by the Wald statistic by accounting for within-subject correlation. The gene-specific score was adjusted by a small positive number to comprise small gene expression level. Permutation test was then performed to compute the false discovery rate (FDR) for each gene. Storey et al. (2005) used a mixed model with a polynomial mean function to detect significant genes across time between treatment and control groups. Under the null hypothesis of no differential expression, the two groups were assumed to have the same population average time curve. The population mean curve for the profile of each gene was modeled, and it was used for calculating F statistic for the gene. And then FDR was used to call significantly differentiated genes.

All methods above consider within-subject correlations, and they are targeted to test

individual genes. However, the following FDR adjustment leads to a high P value for the multiple tests at one hand, and FDR is so conservative that it excludes many positive signals at the other hand (Storey and Tibshirani (2003)). It has become a current trend to test a set of genes simultaneously instead of performing tests for individual genes. A set of genes, selected from biological knowledge from pathway information or literature mining, are tested for variation as a group. One of such knowledge-based approaches was recently reported as Gene set enrichment (Subramanian et al. (2005); Efron and Tibshirani (2007)). Tsai and Qu (2008) tested a subset of genes by applying non-parametric time-varying coefficient model. The within-subject correlation was taken into account by the quadratic inference function (QIF). QIF is derived from GEE and it is asymptotically χ^2 distributed when the number of replications goes to ∞.

Due to the high cost of microarray experiments, a large sample size is usually difficult to obtain. Therefore, the methods based on large sample size have limited application in array study. In addition, the large number of variables and multiple time points entail high requirement for computation. Efficient computation algorithm need to be implemented for methodology development. In microarray data analysis, the raw data are to be pre-processed for quality control and data normalization. In many studies, it is convenient to make normal or log-normal distribution assumption about the raw or processed data (Tseng et al. (2001); Olshen and Jain (2002); Sidorov et al. (2002)). Hoyle et al. (2002) justified that microarray data are in agreement with both Benford's law and Zipf's law, and suggested the lognormal model to be a good candidate concerning the data distribution. However, there are a number of arguments that the data are largely skewed, and the normal or log-normal distributions does not provide a close fit to the data (Kerr et al. (2000); Konishi (2004)). Therefore, a statistical method that has wide application in microarray data analysis should be robust for multiple distribution assumptions and potential outliers.

For Affymetrix microarray chips, the raw microarray data are generated in a similar way as aCGH data described in section 2.2. Both expression microarray and aCGH are

based on fluorescent intensities, which are then transformed to discrimination scores by summarizing the relative measures of perfectly matched mismatched probes. In contrast to aCGH, microarray takes the discrimination scores as raw input data, instead of calculating a ratio to a reference sample, and microarray does not summarize the allele information.

The goal of this chapter is to provide a series of hypothesis testing theory to compare the expression levels for the effect of a set of probes or genes, the time effect and probe by time interactions in a longitudinal microarray study. We would like to use a general model set up so that the test statistics are robust with respect to non-normality. They can also be used for other high dimensional low sample size data with within-subject correlations. The proposed test statistics consider unbalanced designs and heteroscedastic covariance structures as well. An unbalanced design is very common in current microarray data analysis. The data are often collected from different sources such as multiple centers or online database. The dataset often contain different versions or even different manufacturers of microarray. Thus, the number of measurements varies between genes. Our proposed methods can be adapted to various designs. Furthermore, they have the potential to be used with test-based clustering to identify groups of genes with similar expression patterns, producing a gene expression signature.

The outline of this chapter is as follows. In section (3.2), we describe the study design and the model specification. Test statistics are provided in section (3.3). Details of asymptotic theory for original observations and their proofs are provided in section (3.4). Section (3.5) presents the simulation results on type I error estimates and power analysis for our proposed methods. In section (3.6), we applied our method to a recent longitudinal microarray study in which the gene expression profiles of murine T cells with or without interleukin-2 (IL-2) stimulation were collected at 4 and 8 h. Sets of genes from different functional groups were tested for IL-2 signaling over time.

3.2 Model specification

We consider high dimensional longitudinal data in this manuscript. The subjects are randomly assigned to different treatment groups, each subject has thousands of variables, and they are repeatedly measured over time. We focus on the applications of analysis of biological data, such as genomic, proteomics, and metabonomics data.

Let X_{ijkl} be measurement of the kth gene/probe from subject l in treatment group i at time j ($i = 1, ..., I$; $j = 1, ..., J$; $k = 1, ..., K$; $l = 1, ..., n_{ik}$). The number of probes is large, whereas the number of time points and the number of replications are small. The design is assumed to be either balanced or unbalanced, in that the number of replications may vary for different treatment groups and for different probes. The measurement values are modeled by

$$X_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijkl} \tag{3.2.1}$$

where $\sum_i^I \alpha_i = \sum_j^J \beta_j = \sum_k^K \gamma_k = \sum_i^I (\alpha\beta)_{ij} = \sum_j^J (\alpha\beta)_{ij} = \sum_i^I (\alpha\gamma)_{ik} = \sum_k^K (\alpha\gamma)_{ik} = \sum_j^J (\beta\gamma)_{jk} = \sum_k^K (\beta\gamma)_{jk} = 0$, μ is the overall mean measurement, α_i represents the effect of ith treatment, β_j represents the effect of the jth time point, and γ_k represents the effect of the kth gene. The interaction effect of treatment and gene is denoted by $(\alpha\gamma)_{ik}$, and the interaction effect of time and gene is denoted by $(\beta\gamma)_{jk}$. The error terms ε_{ijkl} are identically distributed with mean 0. Assume that ε_{ijkl} and $\varepsilon_{i'j'k'l'}$ are independent if $i \neq i'$ or $k \neq k'$ or $l \neq l'$. They are only dependent when they are observations at various time points for the same gene from an individual, in this case, i=i', $j \neq j'$, k=k', and l=l'. The three-way interaction of treatment, time, and response is not of biological interest, so it is not included in the model. Note that normality is not assumed for ε_{ijkl}. We only requires the existence of the fourth moment.

The treatment effect α_i accounts for differences between treatments averaged over the

whole time period and over all genes. Such differences could arise if the mRNA transcription of some genes is inhibited by the treatment. Similarly, the time effect β_j accounts for the differences between time points. The gene expression may have a trend over time. The gene effect γ_k accounts for the average relative expression level of gene k. The term $(\alpha\gamma)_{ik}$ accounts for the effect of treatment i for gene k. An individual gene could have distinct responses to different treatments. Nonzero treatment*gene interaction indicates differential expression for some genes. The term $(\beta\gamma)_{jk}$ accounts for the effect of gene k at time point j. Genes could have distinct expression trends over time.

The dependence of repeated measurements within an individual was taken into account in the model fitting procedure. The within-subject correlation structure is not necessarily homogeneous, but could vary for different genes. In microarray data, each individual gene has its own transcription activity, therefore, each gene has its unique correlation structure. Biological time course experiments are often not evenly spaced in time. Thus, the same correlation structure should not be assumed for different time point for the same subject. Therefore a heteroscedastic variance structure is used for the model such that $Cov(\varepsilon_{ijkl}, \varepsilon_{ij'kl}) = \sigma_{i,k,jj'}$.

The tests can be written in terms of the parameters in the model. In microarray experiments, the null hypothesis of no differential expression between treatments is equivalent to restricting all α_i to be zero. The test of the null hypothesis that the gene expression does not vary over time is equivalent to testing all β_j equal to zero. The null hypothesis of no variation in gene expression levels between treatments is equivalent to restricting all $(\alpha\gamma)_{ik}$ to zero. The null hypothesis of gene expression independent of time points is equivalent to restricting all $(\beta\gamma)_{jk}$ to zero. In order to identify differentially expressed genes, we will test whether all γ_k equal to zero across all treatments and over the whole time period.

At the end of the section, we present a summary of notations which will be used in the

rest of the manuscript.

$$\widetilde{X}_{i\cdot k\cdot} = J^{-1}\sum_{j=1}^{J}\overline{X}_{ijk\cdot},\ \widetilde{X}_{ij\cdot\cdot} = K^{-1}\sum_{k=1}^{K}\overline{X}_{ijk\cdot},\ \widetilde{X}_{\cdot jk\cdot} = I^{-1}\sum_{i=1}^{I}\overline{X}_{ijk\cdot},\ \widetilde{X}_{i\cdots} = J^{-1}\sum_{j=1}^{J}\widetilde{X}_{ij\cdot\cdot},$$

$$\widetilde{X}_{\cdot j\cdot\cdot} = I^{-1}\sum_{i=1}^{I}\widetilde{X}_{ij\cdot\cdot},\ \sigma^2_{i,k,j} = Var(X_{ijkl}),\ \sigma_{i,k,jj_1} = Cov(X_{ijkl}, X_{ij_1kl})\ (\text{ note } \sigma_{i,k,jj} = \sigma^2_{i,k,j}),$$

$$\sigma^2_{i,k,jj_1} = Var(X_{ijkl}X_{ij_1kl}),\ \sigma_{i,k,jj_1,j_2j_3} = Cov(X_{ijkl}X_{ij_1kl}, X_{ij_2kl}X_{ij_3kl}),\ (\sigma_{i,k,jj_1,jj_1} = \sigma^2_{i,k,jj_1}).$$

3.3 Test statistics

The analysis of variance (ANOVA) is often used for the model specified in last section (Kerr 2000). However, the asymptotic results for traditional ANOVA are not satisfied because the sample size is small and the data may not be normally distributed. Here we will construct new test statistics that are suitable when there is unknown within-subject correlation in the presence of a large number of variables. In this section, we will use a few modified Wald test statistics and modified F test statistics to provide robust tests for main effect and interactions.

First, we will test the treatment main effect. One of the major purposes of a microarray study is identifying changes in expression across various biological conditions, such as different tissues, species, or drug response states. Under the null hypothesis for microarray studies, there is no differential expression of genes between treatments. The null hypothesis is

$$H_0(A):\ all\ \alpha_i = 0,\ for\ i = 1, ..., I.$$

In order to test $H_0(A)$ of the treatment effect, we consider a more general hypothesis $H_0(A_G) : L\alpha = \mathbf{0}$ for a contrast among α_i, where L is a $p \times I$ contrast matrix, $\alpha = (\alpha_1, ..., \alpha_I)'$, and $\mathbf{0}$ is a p dimensional zero vector. If we test the treatment effect for each individual gene, a Wald-type test statistic with estimated correlation structure for the gene may be consid-

ered (Kent 1982). The Wald statistic for the kth gene is

$$W_{A,k} = D'_{A,k}L'(L\widehat{V_{Ak}}L')^{-1}LD_{A,k}$$

where $D_{A,k} = (\widetilde{X}_{1\cdot k\cdot}, ..., \widetilde{X}_{I\cdot k\cdot})'$, and $\widehat{V_{Ak}}$ is the estimated $I \times I$ variance matrix for vector $D_{A,k}$, which is a diagonal matrix for subjects are independent between treatment groups.

The purpose is to perform multivariate test for the treatment effect that takes into account the large number of variables and the within-subject dependence. To adapt the Wald statistic to high dimensional and non-normal data, we proposed a modified Wald-type test statistic for null hypothesis $H_0(A_G)$ that takes into account all genes.

$$W_A = D'_A L'(L\widehat{V_A}L')^{-1}LD_A, \tag{3.3.1}$$

where $D_A = (\widetilde{X}_{1\cdots}, ..., \widetilde{X}_{I\cdots})'$, and $\widehat{V_A} = diag(\widehat{\eta_{A1}}, ..., \widehat{\eta_{AI}})$, which is the estimated covariance matrix for D_A. The term $\widehat{\eta_{Ai}}$ represents the estimation of variance of $\widetilde{X}_{i\cdots}$, and it is defined as

$$\widehat{\eta_{Ai}} = \frac{1}{J^2K^2}\sum_{j_1,j_2}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(X_{ij_1kl} - \overline{X}_{ij_1k\cdot})(X_{ij_2kl} - \overline{X}_{ij_2k\cdot}). \tag{3.3.2}$$

Secondly, we carry out a statistical test for the time effect. In a longitudinal microarray study, researchers are often interested in genes whose expression changes over time, such as cell cycle genes and HOX genes involved in tissue development. The test of the time effect is only intended for a subset of genes of interest, selected from pathways or from a biological database.

The null hypothesis of no time effect is

$$H_0(B): \ all\ \beta_j = 0,\ for\ j = 1, ..., J.$$

Similar to testing $H_0(A)$, in order to test $H_0(B)$ of the time effect, we also consider a more general hypothesis $H_0(B_G) : L\beta = \mathbf{0}$ where L is a $p \times J$ matrix, $\beta = (\beta_1, ..., \beta_J)'$, and

0 is a p dimensional zero vector. A modified Wald-type test statistic is used for testing $H_0(B_G)$.

$$W_B = D_B' L' (L\widehat{V_B} L')^{-1} L D_B, \tag{3.3.3}$$

where $D_B = (\widetilde{X}_{\cdot 1 \cdot \cdot}, ..., \widetilde{X}_{\cdot J \cdot \cdot})'$, and $\widehat{V_B}$ is the estimated $J \times J$ covariance matrix for D_B, with the value at the j_1th row and the j_2th column being

$$\widehat{\eta_B}_{j_1 j_2} = \frac{1}{I^2 K^2} \sum_i^I \sum_k^K \frac{1}{n_{ik}(n_{ik}-1)} \sum_l^{n_{ik}} (X_{ij_1kl} - \overline{X}_{ij_1k\cdot})(X_{ij_2kl} - \overline{X}_{ij_2k\cdot}). \tag{3.3.4}$$

A third hypothesis test is conducted for the main effect on the gene of interest. While the expression levels vary greatly between genes, it is not desirable to compare the expressions directly. To make comparisons between genes, the data should be adjusted to other data, such as a reference or a control dataset, or an alternative dye intensity. In such cases, we often use log-ratio to transform the original gene expression profile. We are interested in detecting discordance of expression pattern between large groups of genes. The null hypothesis is

$$H_0(G): \ all \ \gamma_k = 0, \ for \ k = 1, ..., K.$$

To test $H_0(G)$, similar to analysis of variance, a modified F test statistic is considered.

$$F(G) = \frac{MST_G}{MSE_G}. \tag{3.3.5}$$

But these versions of MST and MSE are slightly different from that of ANOVA, in that

$$MST_G = \frac{IJ}{K-1} \sum_{k=1}^K (\widetilde{X}_{\cdot\cdot k \cdot} - \widetilde{X}_{\cdot\cdot\cdot\cdot})^2 \tag{3.3.6}$$

where $\widetilde{X}_{\cdot\cdot k\cdot}$ is the sample average of all $\widetilde{X}_{i\cdot k\cdot}$ for i=1, ..., I, and in the same way, $\widetilde{X}_{i\cdot k\cdot}$ is the sample average of $\overline{X}_{ijk\cdot}$. $\widetilde{X}_{\cdots\cdot}$ is the sample average of all $\widetilde{X}_{\cdot\cdot k\cdot}$.

$$MSE_G = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l=1}^{n_{ik}}(X_{ijkl}-\overline{X}_{ijk\cdot})(X_{ij_1kl}-\overline{X}_{ij_1k\cdot}). \tag{3.3.7}$$

The definition of MSE_G is different from that of the traditional MSE in that the within-subject correlation over time is taken into account.

The fourth test statistic is for interaction effect of treatment and time. In such situation, we want to identify the gene sets that are activated by treatment in only some specific time points. The expression of genes are affected by treatment, but the effects are only observed after a period of time.

The null hypothesis of no interaction of treatment and time is

$$H_0(AB):\ all\ (\alpha\beta)_{ij}=0,\ for\ i=1,...,I,\ and\ j=1,...,J.$$

Similar to test $H_0(B)$, in order to test $H_0(AB)$ of the time effect, we also consider a more general hypothesis $H_0(AB_G): L(\alpha\beta)=\mathbf{0}$ where L is matrix with p rows and $I\times J$ columns, $(\alpha\beta)$ is the vector of $(\alpha\beta)_{ij}$ with length $I\times J$, and $\mathbf{0}$ is a p dimensional zero vector. A modified Wald-type test statistic is used for testing $H_0(AB_G)$.

$$W_{AB} = D'_{AB}L'(L\widehat{V_{AB}}L')^{-1}LD_{AB}, \tag{3.3.8}$$

where $D_{AB}=(\widetilde{X}_{11\cdot\cdot},\widetilde{X}_{12\cdot\cdot},...,\widetilde{X}_{ij\cdot\cdot},...,\widetilde{X}_{IJ\cdot\cdot})'$, and $\widehat{V_{AB}}$ is the estimated covariance matrix for D_{AB}. The estimated covariance of $\widetilde{X}_{ij\cdot\cdot}$ and $\widetilde{X}_{i_1j_1\cdot\cdot}$ is given at the $((i-1)J+j)$th row and $((i_1-1)J+j_1)$th column of $\widehat{V_{AB}}$. If $i\neq i_1$, the value is zero. If $i=i_1$, the value is given by

$$\widehat{\eta_{AB}}_{(ij)(ij_1)} = \frac{1}{K^2}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(X_{ij_1kl}-\overline{X}_{ij_1k\cdot})(X_{ij_2kl}-\overline{X}_{ij_2k\cdot}). \tag{3.3.9}$$

The next test is targeted to find the gene set that have some genes responsible to the treatment. The gene set can be discovered via hypothesis test of the interaction of gene and treatment effect. The null hypothesis is

$$H_0(AG): \ all \ (\alpha\gamma)_{ik} = 0, \ for \ i = 1, ..., I, \ and \ k = 1, ..., K.$$

Similar to the test statistic for main effect of gene, the test of interaction $H_0(AG)$ is also based on a derivative of F statistic

$$F(AG) = \frac{MST_{AG}}{MSE_{AG}}, \tag{3.3.10}$$

where

$$MST_{AG} = \frac{J}{(I-1)(K-1)} \sum_{i}^{I} \sum_{k}^{K} (\widetilde{X}_{i \cdot k \cdot} - \widetilde{X}_{i \cdots} - \widetilde{X}_{\cdot \cdot k \cdot} + \widetilde{X}_{\cdots \cdot})^2, \tag{3.3.11}$$

$$MSE_{AG} \ = \ \frac{1}{IJK} \sum_{i}^{I} \sum_{j,j_1}^{J} \sum_{k}^{K} \frac{1}{n_{ik}(n_{ik}-1)} \sum_{l}^{n_{ik}} (X_{ijkl} - \overline{X}_{ijk \cdot})(X_{ij_1kl} - \overline{X}_{ij_1k \cdot}).$$

The calculation of sample average is different from ANOVA model, and it is denoted by $\widetilde{X}_{\cdots \cdot})$.

Finally, the high throughput time course study is often targeted to identify the variables that show positive time response, such as genes regulated by cell cycle. We provide hypothesis test for the interaction of variables and time. The null hypothesis is

$$H_0(BG): \ all \ (\beta\gamma)_{jk} = 0, \ for \ j = 1, ..., J, \ and \ k = 1, ..., K.$$

Alike to the test statistic for main effect of gene, the test of interaction $H_0(BG)$ is also based on a derivative of F statistic

$$F(BG) = \frac{MST_{BG}}{MSE_{BG}}, \tag{3.3.12}$$

where

$$MST_{BG} = \frac{I}{(J-1)(K-1)} \sum_{j}^{J} \sum_{k}^{K} (\widetilde{X}_{\cdot jk\cdot} - \widetilde{X}_{\cdot j\cdot\cdot} - \widetilde{X}_{\cdot\cdot k\cdot} + \widetilde{X}_{\cdot\cdot\cdot\cdot})^2, \tag{3.3.13}$$

$$\begin{aligned} MSE_{BG} &= \frac{1}{IK(J-1)} \sum_{i}^{I} \sum_{j}^{J} \sum_{k}^{K} \frac{1}{n_{ik}(n_{ik}-1)} \sum_{l}^{n_{ik}} (X_{ijkl} - \overline{X}_{ijk\cdot})^2 - \\ & \frac{1}{IKJ(J-1)} \sum_{i}^{I} \sum_{j,j_1}^{J} \sum_{k}^{K} \frac{1}{n_{ik}(n_{ik}-1)} \sum_{l}^{n_{ik}} (X_{ijkl} - \overline{X}_{ijk\cdot})(X_{ij_1kl} - \overline{X}_{ij_1k\cdot}). \end{aligned}$$

The calculation of sample average is different from ANOVA model, and it is denoted by $\widetilde{X}_{\cdot\cdot\cdot\cdot}$.

The asymptotic distribution for each of the test statistic will be derived in the following sections.

3.4 Main results based on original observations

This section is devoted to develop the asymptotic distribution of the test statistics which are defined in the last section. The asymptotic properties are derived based on the original observation under null hypotheses. For simplicity, we use the residual $e_{ijkl} = X_{ijkl} - E[X_{ijkl}]$ in this section.

Theorem 3.4.1. *For testing $H_0(A_G)$: $L\alpha = \mathbf{0}$ where L is a $I \times p$ contrast matrix, $\alpha = (\alpha_1, ..., \alpha_I)'$, and $\mathbf{0}$ is a p dimensional zero vector, let W_A be the statistic given in (3.3.1). If X_{ijkl} has a finite fourth moment, then under $H_0(A_G)$,*

$$W_A \xrightarrow{d} \chi^2_p \text{ as } K \to \infty.$$

We will start the proof by first showing that the variance estimation in W_A is consistent.

Lemma 3.4.2. *Let $\eta_{A,i} = Var(\widetilde{X}_{i\cdots})$ denote the variance of $\widetilde{X}_{i\cdots}$, and let $\widehat{\eta_{A,i}}$ be the statistic given in (3.3.2). Under the settings and assumptions of Theorem 3.4.1,*

$$K(\widehat{\eta_{A,i}} - \eta_{A,i}) \xrightarrow{p} 0 \text{ as } K \to \infty.$$

for i=1, ..., I.

Proof: By the independence of $\widetilde{X}_{i\cdot k\cdot}$ for k=1, ..., K, we have

$$\eta_{A,i} = Var(\widetilde{X}_{i\cdots}) = Var\left(\frac{\sum_{k=1}^{K}\widetilde{X}_{i\cdot k\cdot}}{K}\right) = \frac{1}{K^2}\sum_{k=1}^{K} Var(\widetilde{X}_{i\cdot k\cdot}).$$

Let

$$\widehat{\eta_{Ai,k}} = \frac{1}{J^2 n_{ik}(n_{ik}-1)}\sum_{j_1,j_2}^{J}\sum_{l}^{n_{ik}}(X_{ij_1kl} - \overline{X}_{ij_1k\cdot})(X_{ij_2kl} - \overline{X}_{ij_2k\cdot}).$$

We will show that $\widehat{\eta_{Ai,k}}$ is an unbiased estimator of $\eta_{A,i,k} = Var(\widetilde{X}_{i\cdot k\cdot})$. First note that

$$\begin{aligned}
& E\left[(X_{ij_1kl} - \overline{X}_{ij_1k\cdot})(X_{ij_2kl} - \overline{X}_{ij_2k\cdot})\right] \\
= \; & E\left[(e_{ij_1kl} - \overline{e}_{ij_1k\cdot})(e_{ij_2kl} - \overline{e}_{ij_2k\cdot})\right] \\
= \; & E(e_{ij_1kl}e_{ij_2kl}) - E(\overline{e}_{ij_1k\cdot}e_{ij_2kl}) - E(e_{ij_1kl}\overline{e}_{ij_2k\cdot}) + E(\overline{e}_{ij_1k\cdot}\overline{e}_{ij_2k\cdot}) \\
= \; & \sigma_{i,k,j_1j_2} - \frac{1}{n_{ik}}\sigma_{i,k,j_1j_2} - \frac{1}{n_{ik}}\sigma_{i,k,j_1j_2} + \frac{1}{n_{ik}^2}\sum_{l}^{n_{ik}} E(e_{ij_1kl}e_{ij_2kl}) \\
= \; & \frac{n_{ik}-1}{n_{ik}}\sigma_{i,k,j_1j_2}.
\end{aligned}$$

We then have

$$E(\widehat{\eta_{Ai,k}}) = \frac{1}{J^2 n_{ik}(n_{ik}-1)}\sum_{j_1,j_2}^{J}\sum_{l}^{n_{ik}} E\left[(X_{ij_1kl} - \overline{X}_{ij_1k\cdot})(X_{ij_2kl} - \overline{X}_{ij_2k\cdot})\right] = \frac{1}{J^2 n_{i,k}}\sum_{j_1j_2}^{J}\sigma_{i,k,j_1j_2}$$

It is easy to show that

$$\eta_{A,i,k} = \frac{1}{J^2 n_{i,k}}\sum_{j_1j_2}^{J}\sigma_{i,k,j_1j_2}.$$

Therefore we have shown that $E(\widehat{\eta_{A_{i,k}}}) = \eta_{A,i,k}$. The lemma will follow by showing that $\frac{1}{K}\sum_{k=1}^{K}(\widehat{\eta_{A_{i,k}}} - \eta_{A,i,k}) \xrightarrow{p} 0$. The convergence is obtained by applying the Markov weak law of large number. The Markov condition will be satisfied if $\frac{1}{K^2}\sum_{k=1}^{K} E(\widehat{\eta_{A_{i,k}}} - \eta_{A,i,k})^2 \to 0$ as $n \to \infty$. It is sufficient to show that $E(\widehat{\eta_{A_{i,k}}})^2$ is finite. By Hölder's inequality,

$$
\begin{aligned}
& E(\widehat{\eta_{A_{i,k}}})^2 \\
= \; & E\left[\frac{1}{J^2 n_{ik}(n_{ik}-1)}\sum_{j_1,j_2}^{J}\sum_{l}^{n_{ik}}(e_{ij_1kl} - \overline{e}_{ij_1k\cdot})(e_{ij_2kl} - \overline{e}_{ij_2k\cdot})\right]^2 \\
= \; & \frac{1}{J^4 n_{ik}^2(n_{ik}-1)^2} E\left[\sum_{j_1 J_2}^{J}\sum_{l}^{n_{ik}}(e_{ij_1kl} - \overline{e}_{ij_1k\cdot})(e_{ij_2kl} - \overline{e}_{ij_2k\cdot})\right]^2 \\
\leq \; & \frac{1}{J^2 n_{ik}(n_{ik}-1)^2}\sum_{j_1 J_2}^{J}\sum_{l}^{n_{ik}} E\left[(e_{ij_1kl} - \overline{e}_{ij_1k\cdot})(e_{ij_2kl} - \overline{e}_{ij_2k\cdot})\right]^2 \\
= \; & \frac{1}{J^2 n_{ik}(n_{ik}-1)^2}\sum_{j_1 J_2}^{J}\sum_{l}^{n_{ik}} E\left[e_{ij_1kl}e_{ij_2kl} - e_{ij_2kl}\overline{e}_{ij_1k\cdot} - e_{ij_1kl}\overline{e}_{ij_1k\cdot} + \overline{e}_{ij_2k\cdot}\overline{e}_{ij_2k\cdot}\right]^2 \\
\leq \; & \frac{4}{J^2 n_{ik}(n_{ik}-1)^2}\sum_{j_1 J_2}^{J}\sum_{l}^{n_{ik}} \Big[E(e_{ij_1kl}e_{ij_2kl})^2 + E(e_{ij_2kl}\overline{e}_{ij_1k\cdot})^2 + E(e_{ij_1kl}\overline{e}_{ij_1k\cdot})^2 + \\
& E(\overline{e}_{ij_2k\cdot}\overline{e}_{ij_2k\cdot})^2\Big] \\
< \; & \infty,
\end{aligned}
$$

for fixed J and n_{ik}. The finite bound is obtained because the first four moments of X_{ijkl} exist. This completes the proof.

Proof of Theorem 3.4.1: Under $H_0(A_G)$, $LE[D_A] = \mathbf{0}$, where $D_A = (\widetilde{X}_{1\cdots}, ..., \widetilde{X}_{I\cdots})'$, we have $LD_A = LD_A - E[D_A])$. Let $V_A = Var[D_A] = diag(\eta_{A,1}, ..., \eta_{A,I})$. From Lemma (3.4.2), we have that $\widehat{V_A}$ is a consistent estimate of V_A. Because of the independence of $\widetilde{X}_{1\cdots}$'s, the result will follow with the Continuous Mapping and Slutsky's Theorems, by showing $(\widetilde{X}_{i\cdots} - E[\widetilde{X}_{i\cdots}])/\sqrt{\eta_{A,i}} \xrightarrow{d} N(0,1)$ as $K \to \infty$. Since $\widetilde{X}_{i\cdot k\cdot}$'s are independent for k=1, ...,K, it is sufficient to show that $\sum_k(\widetilde{X}_{i\cdot k\cdot} - E[\widetilde{X}_{i\cdot k\cdot}])/(K\sqrt{\eta_{A,i}}) = \sum_k \widetilde{e}_{i\cdot k\cdot}/(K\sqrt{\eta_{A,i}}) \xrightarrow{d} N(0,1)$ as $K \to \infty$. And the asymptotic normality of $\widetilde{X}_{i\cdots}$ can be shown by Lyapounov's Theorem.

The Lyapounov condition will be satisfied if

$$L_A = \frac{\sum_{k=1}^{K} E|\widetilde{e}_{i\cdot k\cdot}|^4}{(\sum_{k=1}^{K} \eta_{A,i,k})^2} \to 0$$

Because the first four moments of X_{ijkl} exists, it is easy to show $E|\widetilde{e}_{i\cdot k\cdot}|^4$ is finite for any k by Hölder's theorem. Since $\eta_{A,i,k}$'s are non-zero constant, $L_A \to 0$ as $a \to \infty$. This completes the proof.

Theorem 3.4.3. *For testing $H_0(B_G)$: $L\beta = \mathbf{0}$ where L is a $p \times J$ contrast matrix, $\beta = (\beta_1, ..., \beta_J)'$, and $\mathbf{0}$ is a p dimensional zero vector, let W_B be the statistic given in (3.3.3). If X_{ijkl} has a finite fourth moment, then under $H_0(B_G)$,*

$$W_B \xrightarrow{d} \chi^2_p \text{ as } K \to \infty.$$

Proof of Theorem 3.4.3: Under $H_0(B_G)$, $LE[D_B] = \mathbf{0}$, then $LD_B = L(D_B - E[D_B])$. Let $V_B = Var[D_B]$. V_B is a $J \times J$ matrix, where the value of j_1th row and j_2th column is defined as

$$Cov(\widetilde{X}_{\cdot j_1\cdot\cdot}, \widetilde{X}_{\cdot j_2\cdot\cdot}) = \eta_{B,j_1j_2} = \frac{1}{I^2K^2} \sum_{i}^{I} \sum_{k}^{K} \frac{\sigma_{i,k,j_1j_2}}{n_{ik}}.$$

If $j_1 = j_2 = j$, it is the variance of $\widetilde{X}_{\cdot j\cdot\cdot}$, and it is denoted

$$\eta_{B,j} = \frac{1}{I^2K^2} \sum_{i}^{I} \sum_{k}^{K} \frac{\sigma^2_{i,k,j}}{n_{ik}}.$$

The result will follow with the Continuous Mapping and Slutsky's Theorems, by showing $\sqrt{N}(D_B - E[D_B]) \xrightarrow{d} N_J(\mathbf{0}, \lim_{K\to\infty} NV_B)$ as $K \to \infty$, where $N = \sum_{i=1}^{I} \sum_{k=1}^{K} n_{ik}$. It is sufficient to show for any finite constants $\mathbf{a} = (a_1, ..., a_J)'$,

$$\sqrt{N}\mathbf{a}'(D_B - E[D_B]) = \sum_{j=1}^{J} a_j \sqrt{N}(\widetilde{X}_{\cdot j\cdot\cdot} - E(\widetilde{X}_{\cdot j\cdot\cdot})) \xrightarrow{d} N(0, \lim_{K\to\infty} N\mathbf{a}'V_B\mathbf{a}),$$

where $N\mathbf{a}'V_B\mathbf{a} = N \sum_{i=1}^{I} \sum_{j,j_1}^{J} \sum_{k}^{K} a_j a_{j_1} \sigma_{i,k,jj_1}/(n_{ik}I^2K^2)$ converges if $(\sum_i^I \sum_k^K n_{ik}^{-1})(\sum_i^I \sum_k^K n_{ik})/K^2$ converges, which holds if $\max n_{ik}/\min n_{ik} = O(1)$.

Note that

$$\begin{aligned}
&\sum_{j=1}^{J} a_j \sqrt{N}[\widetilde{X}_{\cdot j\cdot} - E(\widetilde{X}_{\cdot j\cdot})] \\
= \quad & \frac{\sqrt{N}}{K} \sum_{j=1}^{J} a_j \sum_{k}^{K} [\widetilde{X}_{\cdot jk\cdot} - E(\widetilde{X}_{\cdot jk\cdot})] \\
= \quad & \sum_{k}^{K} \left(\frac{\sqrt{N}}{K} \sum_{j}^{J} a_j \widetilde{e}_{\cdot jk\cdot} \right).
\end{aligned}$$

Asymptotic normality is attained by applying Lyapounov's Theorem. The Lyapounov condition will be satisfied if

$$L_B = \sum_{k}^{K} \left(\frac{\sqrt{N}}{K} \right)^4 E \left(\sum_{j}^{J} a_j \widetilde{e}_{\cdot jk\cdot} \right)^4 \to 0.$$

To see this,

$$\begin{aligned}
L_B \quad = \quad & \sum_{k}^{K} \left(\frac{\sqrt{N}}{K} \right)^4 E \left(\sum_{j}^{J} a_j \widetilde{e}_{\cdot jk\cdot} \right)^4 \leq \sum_{k}^{K} \frac{N^2}{K^4} J^3 \sum_{j}^{J} E(a_j^4 \widetilde{e}_{\cdot jk\cdot}^4) \\
= \quad & \sum_{k}^{K} \frac{N^2}{K^4} J^3 \sum_{j}^{J} \frac{a_j^4}{I} E(\sum_{i}^{I} \overline{e}_{ijk\cdot}^4) \leq \sum_{k}^{K} \frac{N^2}{K^4} J^3 \sum_{j}^{J} \frac{a_j^4}{I} B_4 \left[\sum_{i}^{I} E(\overline{e}_{ijk\cdot}^2) \right]^2 \\
= \quad & \sum_{k}^{K} \frac{N^2}{K^4} J^3 \sum_{j}^{J} \frac{a_j^4}{I} B_4 \left[\sum_{i}^{I} \frac{1}{n_{ik}^2} E(\sum_{l}^{n_{ik}} e_{ijkl})^2 \right]^2 \\
= \quad & \sum_{k}^{K} \frac{N^2}{K^4} J^3 \sum_{j}^{J} \frac{a_j^4}{I} B_4 \left[\sum_{i}^{I} \frac{1}{n_{ik}} \sigma_{i,k,j}^2) \right]^2 \\
= \quad & O\left(\sum_{k}^{K} \frac{N^2}{K^4} \sum_{i}^{I} \frac{1}{n_{ik}^2} \right) \\
= \quad & O(K^{-1}),
\end{aligned}$$

where the first inequality follows Hölder's inequality (2.5.2), and the last equality holds if $max\{n_{ik}\} = O(min\{n_{ik}\})$. The second inequality follows from Khintchine's inequality (2.5.4). B_4 is a constant with definition of $B_{2m} = ((2m)!/2^m m!)^{\frac{1}{2m}}$.

This completes the proof.

Theorem 3.4.4. *For null hypothesis $H_0(G)$: all $\gamma_k = 0$, let $F(G)$ be the statistic given in (3.3.5). If X_{ijkl} has a finite fourth moment, then under $H_0(G)$, as $K \to \infty$,*

$$\frac{\sqrt{K}(F(G)-1)}{V_G} \xrightarrow{d} N(0,1),$$

where V_G is the asymptotic variance defined as

$$V_G = \sqrt{\tau_G}/\sigma_G. \tag{3.4.1}$$

where

$$\tau_G = \frac{2}{I^2J^2K}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{j,j_1,j_2,j_3}^{J}\frac{1}{n_{ik}n_{i_1k}}\sigma_{i,k,jj_1}\sigma_{i_1,k,j_2j_3} + \sum_{i}^{I}\sum_{j,j_1,j_2,j_3}^{J}\frac{1}{n_{ik}(n_{ik}-1)}\sigma_{i,k,jj_1}\sigma_{i,k,j_2j_3}\right],$$

$$\sigma_G = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{\sigma_{i,k,jj_1}}{n_{ik}}.$$

Lemma 3.4.5. *Under the settings and assumptions of Theorem 3.4.4,*

$$MSE_G - \sigma_G \xrightarrow{p} 0 \text{ as } K \to \infty.$$

Proof:

First note that in the proof of Lemma (3.4.2) we have shown

$$E\left[(X_{ij_1kl} - \overline{X}_{ij_1k\cdot})(X_{ij_2kl} - \overline{X}_{ij_2k\cdot})\right] = \frac{n_{ik}-1}{n_{ik}}\sigma_{i,k,j_1j_2}.$$

Then,

$$\begin{aligned} E(MSE_G) &= \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l=1}^{n_{ik}}E\left[(X_{ijkl} - \overline{X}_{ijk\cdot})(X_{ij_1kl} - \overline{X}_{ij_1k\cdot})\right] \\ &= \sigma_G. \end{aligned}$$

And we have

$$
\begin{aligned}
& Var(MSE_G) \\
= \; & \frac{1}{(IJK)^2} \sum_{i=1}^{I} \sum_{j,j_1}^{J} \sum_{k} \frac{1}{n_{ik}^2 (n_{ik}-1)^2} \sum_{l=1}^{n_{ik}} \sum_{l_2=1}^{n_{ik}} Cov \left[\sum_{j,j_1}^{b} (X_{ijkl} - \overline{X}_{ij.})(X_{ij_1kl} - \overline{X}_{ij_1.}), \right. \\
& \left. \sum_{j_2,j_3}^{b} (X_{ij_2kl_2} - \overline{X}_{ij_2.})(X_{ij_3kl_2} - \overline{X}_{ij_3.}) \right].
\end{aligned}
$$

Note that

$$
\begin{aligned}
& E\left[(X_{ijkl} - \overline{X}_{ijk.})(X_{ij_1kl} - \overline{X}_{ij_1k.})\right]^2 \\
= \; & E\left[(e_{ijkl} - \overline{e}_{ijk.})(e_{ij_1kl} - \overline{e}_{ij_1k.})\right]^2 \\
= \; & E[e_{ijkl}e_{ij_1kl} - \overline{e}_{ijk.}e_{ij_1kl} - e_{ijkl}\overline{e}_{ij_1k.} + \overline{e}_{ijk.}\overline{e}_{ij_1k.}]^2 \\
\leq \; & 4[E(e_{ijkl}e_{ij_1kl})^2 + E(\overline{e}_{ijk.}e_{ij_1kl})^2 + E(e_{ijkl}\overline{e}_{ij_1k.})^2 + E(\overline{e}_{ijk.}\overline{e}_{ij_1k.})^2] \\
= \; & 4\left[E(e_{ijkl}^2 e_{ij_1kl}^2) + \frac{1}{n_{ik}^2}E(e_{ijkl}^2 e_{ij_1kl}^2) + \frac{1}{n_{ik}^2}E(e_{ijkl}^2 e_{ij_1kl}^2) + \frac{n_{ik}^2}{n_{ik}^4}\sum_{l=1}^{n_{ik}}\sum_{l_1=1}^{n_{ik}} E(e_{ijkl}^2 e_{ij_1kl_1}^2)\right] \\
= \; & \frac{4(n_{ik}^2 + 2 + n_{ik})}{n_{ik}^2} Cov(e_{ijkl}^2, e_{ij_1kl}^2) + 4\sigma_{ijk}^2\sigma_{ij_1k}^2 \\
< \; & \infty,
\end{aligned}
$$

where the first inequality follows Hölder's inequality (2.5.2), and the last inequality holds because X_{ijk} has the finite fourth moment.

We have

$$\left|Cov\left[\sum_{j,j_1}^{J}(e_{ijkl}-\overline{e}_{ijk.})(e_{ij_1kl}-\overline{e}_{ij_1k.}),\sum_{j_2,j_3}^{J}(e_{ij_2kl_2}-\overline{e}_{ij_2k.})(e_{ij_3kl_2}-\overline{e}_{ij_3k.})\right]\right| \tag{3.4.2}$$

$$\leq \left|Var\left[\sum_{j,j_1}^{J}(e_{ijkl}-\overline{e}_{ijk.})(e_{ij_1kl}-\overline{e}_{ij_1k.})\right]Var\left[\sum_{j_2,j_3}^{J}(e_{ij_2kl_2}-\overline{e}_{ij_2k.})(e_{ij_3kl_2}-\overline{e}_{ij_3k.})\right]\right|^{\frac{1}{2}}$$

$$\leq \left|E\left[\sum_{j,j_1}^{J}(e_{ijkl}-\overline{e}_{ijk.})(e_{ij_1kl}-\overline{e}_{ij_1k.})\right]^2\right|^{\frac{1}{2}}\left|E\left[\sum_{j_2,j_3}^{J}(e_{ij_2kl_2}-\overline{e}_{ij_2k.})(e_{ij_3kl_2}-\overline{e}_{ij_3k.})\right]^2\right|^{\frac{1}{2}}$$

$$\leq \left|J^2\sum_{j,j_1}^{J}E\left[(e_{ijkl}-\overline{e}_{ijk.})(e_{ij_1kl}-\overline{e}_{ij_1k.})\right]^2\right|^{\frac{1}{2}}\left|J^2\sum_{j_2,j_3}^{J}E\left[(e_{ij_2kl_2}-\overline{e}_{ij_2k.})(e_{ij_3kl_2}-\overline{e}_{ij_3k.})\right]^2\right|^{\frac{1}{2}}$$

$$= J^2\left|\sum_{j,j_1}^{J}E\left[(e_{ijkl}-\overline{e}_{ijk.})(e_{ij_1kl}-\overline{e}_{ij_1k.})\right]^2\right|^{\frac{1}{2}}\left|\sum_{j_2,j_3}^{J}E\left[(e_{ij_2kl_2}-\overline{e}_{ij_2k.})(e_{ij_3kl_2}-\overline{e}_{ij_3k.})\right]^2\right|^{\frac{1}{2}}$$

$$< \infty,$$

where the inequalities follow from the Cauchy Schwartz Inequality and Hölder's inequality, and the last inequality holds due to the results previously shown.

Therefore,

$$Var(MSE_G) = O(K^{-1})$$

It follows that $MSE_G - \sigma_G \xrightarrow{p} 0$ as $K \to \infty$.

Lemma 3.4.6. *Under the settings and assumptions of Theorem 3.4.4 and under $H_0(G)$, we have*

$$\sqrt{K}(MST_G - P_G(e)) \xrightarrow{p} 0 \text{ as } K \to \infty,$$

where $P_G(e) = \frac{IJ}{K}\sum_{k=1}^{K}\widetilde{e}^2_{\cdot\cdot k\cdot}$.

Proof:

Note that under $H_0(G)$,

$$\begin{aligned} MST_G &= \frac{IJ}{K-1}\sum_{k=1}^{K}(\widetilde{e}_{\cdot\cdot k\cdot} - \widetilde{e}_{\cdot\cdot\cdot\cdot})^2 \\ &= \frac{IJ}{K-1}\left(\frac{K-1}{K}\sum_{k=1}^{K}\widetilde{e}^2_{\cdot\cdot k\cdot} - \frac{1}{K}\sum_{k\neq k'}^{K}\widetilde{e}_{\cdot\cdot k\cdot}\widetilde{e}_{\cdot\cdot k'\cdot}\right) \\ &= \frac{IJ}{K}\sum_{k=1}^{K}\widetilde{e}^2_{\cdot\cdot k\cdot} - \frac{IJ}{K(K-1)}\sum_{k\neq k'}^{K}\widetilde{e}_{\cdot\cdot k\cdot}\widetilde{e}_{\cdot\cdot k\cdot}. \end{aligned}$$

Thus, we have

$$E[\sqrt{K}(MST_G - P_G(e))] = \frac{\sqrt{K}IJ}{K(K-1)}\sum_{k\neq k'}^{K} E\left[\widetilde{e}_{\cdot\cdot k\cdot}\widetilde{e}_{\cdot\cdot k\cdot}\right] = 0.$$

Furthermore,

$$\begin{aligned} & E[\sqrt{K}(MST_G - P_G(e))]^2 \\ = & \frac{KI^2J^2}{K^2(K-1)^2}\left(\sum_{k\neq k'}^{K} E\left[\widetilde{e}_{\cdot\cdot k\cdot}\widetilde{e}_{\cdot\cdot k\cdot}\right]\right)^2 \\ = & \frac{I^2J^2}{K(K-1)^2}E\left[\sum_{k\neq k_1, k_2\neq k_3}^{K}\widetilde{e}^2_{\cdot\cdot k\cdot}\widetilde{e}^2_{\cdot\cdot k_1\cdot}\widetilde{e}^2_{\cdot\cdot k_2\cdot}\widetilde{e}^2_{\cdot\cdot k_3\cdot}\right] \\ = & \frac{2I^2J^2}{K(K-1)^2}E\left[\sum_{k\neq k_1}^{K}\widetilde{e}^2_{\cdot\cdot k\cdot}\widetilde{e}^2_{\cdot\cdot k_1\cdot}\right] \\ = & \frac{2I^2J^2}{K(K-1)^2}\sum_{k\neq k_1}^{K} E[\widetilde{e}^2_{\cdot\cdot k\cdot}]E[\widetilde{e}^2_{\cdot\cdot k_1\cdot}] \\ = & O(K^{-1}). \end{aligned}$$

Therefore, under $H_0(G)$, $\sqrt{K}(MST_G - P_G(e)) \xrightarrow{p} 0$ as $K \to \infty$.

Proof of Theorem 3.4.4: From Lemmas 3.4.5 and 3.4.6, we need only to consider the asymptotic distribution of $Q_G(e) = \sqrt{K}(P_G(e) - MSE_G)$ under $H_0(G)$, where $P_G(e) = \frac{IJ}{K}\sum_{k=1}^{K}\widetilde{e}^2_{\cdot\cdot k\cdot}$.

Using some simple algebra, we have

$$
\begin{aligned}
& Q_G(e) \\
= \; & \sqrt{K}\left[\frac{IJ}{K}\sum_{k=1}^{K}\widetilde{e}^{2}_{\cdot\cdot k\cdot} - \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l=1}^{n_{ik}}(X_{ijkl}-\overline{X}_{ijk\cdot})(X_{ij_1kl}-\overline{X}_{ij_1k\cdot})\right] \\
= \; & \frac{1}{IJ\sqrt{K}}\sum_{k=1}^{K}\left[(\sum_{i}^{I}\sum_{j}^{J}\sum_{l}^{n_{ik}}\frac{e_{ijkl}}{n_{ik}})^2 - \sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l=1}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})(e_{ij_1kl}-\overline{e}_{ij_1k\cdot})\right] \\
= \; & \frac{1}{IJ\sqrt{K}}\sum_{j,j_1}^{J}\sum_{k=1}^{K}\left[\sum_{i,i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}} - \sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l=1}^{n_{ik}}(e_{ijkl}e_{ij_1kl} - e_{ij_1kl}\overline{e}_{ijk\cdot}\right. \\
& \left. -e_{ijkl}\overline{e}_{ijk\cdot} + \overline{e}_{ijk\cdot}\overline{e}_{ij_1k\cdot})\right] \\
= \; & \frac{1}{IJ\sqrt{K}}\sum_{j,j_1}^{J}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}} + \sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}\right].
\end{aligned}
$$

Therefore, $E[Q_G] = 0$. It follows that

$$
\begin{aligned}
& Var(Q_G(e)) \\
= \; & E\left(\frac{1}{IJ\sqrt{K}}\sum_{j,j_1}^{J}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}} + \sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ij_1kl_1}\right]\right)^2 \\
= \; & \frac{1}{I^2J^2K}\sum_{k=1}^{K} E\left(\sum_{j,j_1}^{J}\left[\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}} + \sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ij_1kl_1}\right]\right)^2 \\
= \; & \frac{1}{I^2J^2K}\sum_{k=1}^{K}\left[E(\sum_{i\neq i_1}^{I}\sum_{j,j_1}^{J}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}})^2 + E(\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ij_1kl_1})^2\right] \\
= \; & \frac{2}{I^2J^2K}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\frac{1}{n_{ik}^2n_{i_1k}^2}E(\sum_{j,j_1}^{J}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}} e_{ijkl}e_{i_1j_1kl_1})^2 + \sum_{l\neq l_1}^{n_{ik}}E(\sum_{i=1}^{I}\sum_{j,j_1}^{J}\frac{1}{n_{ik}(n_{ik}-1)}e_{ijkl}e_{ij_1kl_1})^2\right] \\
= \; & \frac{2}{I^2J^2K}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{j,j_1,j_2,j_3}^{J}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{n_{ik}^2n_{i_1k}^2}E(e_{ijkl}e_{ij_1kl}e_{i_1j_2kl_1}e_{i_1j_3kl_1})\right. \\
& \left. + \sum_{i}^{I}\sum_{j,j_1,j_2,j_3}^{J}\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}^2(n_{ik}-1)^2}E(e_{ijkl}e_{ij_1kl}e_{ij_2kl_1}e_{ij_3kl_1})\right] \\
= \; & \frac{2}{I^2J^2K}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{j,j_1,j_2,j_3}^{J}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{n_{ik}^2n_{i_1k}^2}E(e_{ijkl}e_{ij_1kl})E(e_{i_1j_2kl_1}e_{i_1j_3kl_1})\right. \\
& \left. + \sum_{i}^{I}\sum_{j,j_1,j_2,j_3}^{J}\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}^2(n_{ik}-1)^2}E(e_{ijkl}e_{ij_1kl})E(e_{ij_2kl_1}e_{ij_3kl_1})\right] \\
= \; & \frac{2}{I^2J^2K}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{j,j_1,j_2,j_3}^{J}\frac{1}{n_{ik}n_{i_1k}}\sigma_{i,k,jj_1}\sigma_{i_1,k,j_2j_3} + \sum_{i}^{I}\sum_{j,j_1,j_2,j_3}^{J}\frac{1}{n_{ik}(n_{ik}-1)}\sigma_{i,k,jj_1}\sigma_{i,k,j_2j_3}\right] \\
= \; & \tau_G.
\end{aligned}
$$

Since $Var(Q_G(e))$ is bounded, Lyapunov's condition will be satisfied if

$$
\begin{aligned}
L_G \; &= \; \sum_{k=1}^{K} E\left|\frac{1}{IJ\sqrt{K}}\sum_{j,j_1}^{J}[\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}} + \sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ij_1kl_1}]\right|^4 \\
&\to \; 0.
\end{aligned}
$$

We have

$$
\begin{aligned}
L_G &= \frac{1}{I^4J^4K^2}\sum_{k=1}^{K}E\left|\sum_{j,j_1}^{J}[\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}}+\sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}]\right|^4 \\
&\leq \frac{J^6}{I^4J^4K^2}\sum_{j,j_1}^{J}\sum_{k=1}^{K}E\left|\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}}+\sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}\right|^4 \\
&\leq \frac{8J^2}{I^4K^2}\sum_{j,j_1}^{J}\sum_{k=1}^{K}\left[E|\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{e_{ijkl}e_{i_1j_1kl_1}}{n_{ik}n_{i_1k}}|^4+E|\sum_{i=1}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}|^4\right] \\
&\leq \frac{8J^2}{I^4K^2}\sum_{j,j_1}^{J}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{I^3(I-1)^3n_{ik}^3n_{i_1k}^3}{n_{ik}^4n_{i_1k}^4}E|e_{ijkl}e_{i_1j_1kl_1}|^4\right. \\
&\quad \left.+\sum_{i=1}^{I}\sum_{l\neq l_1}^{n_{ik}}\frac{I^3n_{ik}^3(n_{ik}-1)^3}{n_{ik}^4(n_{ik}-1)^4}E|e_{ijkl}e_{ij_1kl_1}|^4\right] \\
&= \frac{8J^2}{IK^2}\sum_{j,j_1}^{J}\sum_{k=1}^{K}\left[\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{(I-1)^3}{n_{ik}n_{i_1k}}E(e_{ijkl})^4E(e_{i_1j_1kl_1})^4\right. \\
&\quad \left.+\sum_{i=1}^{I}\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}(n_{ik}-1)}E(e_{ijkl})^4E(e_{ij_1kl_1})^4\right] \\
&= O(K^{-1}),
\end{aligned}
$$

if the fourth moment of e_{ijk} exists for any i, j, and k. The inequalities follow from Hölder's inequality, and the last equation results from the finite fourth central moment of X_{ijkl}. This completes the proof.

Theorem 3.4.7. *For testing $H_0(AB_G)$: $L(\alpha\beta) = \mathbf{0}$ where L is a contrast matrix with p rows and $I \times J$ columns, $(\alpha\beta)$ is the vector of $(\alpha\beta)_{ij}$, and $\mathbf{0}$ is a p dimensional zero vector, let W_{AB} be the statistic given in (3.3.8). If X_{ijkl} has a finite fourth moment, then under $H_0(AB_G)$,*

$$W_{AB} \xrightarrow{d} \chi^2_p \text{ as } K \to \infty.$$

Proof of Theorem 3.4.7: Under $H_0(AB_G)$, $LE[D_{AB}] = \mathbf{0}$, then $LD_{AB} = L(D_{AB} - E[D_{AB}])$. Let $V_{AB} = Var[D_{AB}]$. V_{AB} is a $(IJ) \times (IJ)$ matrix, and the estimated covariance

of $\widetilde{X}_{ij\cdot\cdot}$ and $\widetilde{X}_{i_1j_1\cdot\cdot}$ is given at the $((i-1)J+j)$th row and $((i_1-1)J+j_1)$th column of $\widehat{V_{AB}}$. If $i \neq i_1$, the value is zero. If $i = i_1$, the value is given by

$$\widehat{\eta_{AB}}_{(ij)(ij_1)} = \frac{1}{K^2}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(X_{ij_1kl} - \overline{X}_{ij_1k\cdot})(X_{ij_2kl} - \overline{X}_{ij_2k\cdot}). \tag{3.4.3}$$

The result will follow with the Continuous Mapping and Slutsky's Theorems, by showing $\sqrt{N}(D_{AB} - E[D_{AB}]) \xrightarrow{d} N_J(\mathbf{0}, \lim_{K\to\infty} NV_{AB})$ as $K \to \infty$, where $N = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ik}$. It is sufficient to show for any finite constants $\mathbf{a} = (a_{11}, a_{12}, ..., a_{ij}, ..., a_{IJ})'$,

$$\sqrt{N}\mathbf{a}'(D_{AB} - E[D_{AB}]) = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\sqrt{N}(\widetilde{X}_{ij\cdot\cdot} - E(\widetilde{X}_{ij\cdot\cdot})) \xrightarrow{d} N(0, \lim_{K\to\infty} N\mathbf{a}'V_{AB}\mathbf{a}),$$

where $N\mathbf{a}'V_{AB}\mathbf{a} = N\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K} a_{ij}a_{ij_1}\sigma_{i,k,jj_1}/(n_{ik}K^2)$ converges if $(\sum_i^I\sum_k^K n_{ik}^{-1})(\sum_i^I\sum_k^K n_{ik})/K^2$ converges, which holds if $\max n_{ik}/\min n_{ik} = O(1)$.

$$\begin{aligned}
&\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\sqrt{N}(\widetilde{X}_{ij\cdot\cdot} - E(\widetilde{X}_{ij\cdot\cdot})) \\
= \;& \frac{\sqrt{N}}{K}\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\sum_{k}^{K}(\overline{X}_{ijk\cdot} - E(\overline{X}_{ijk\cdot})) \\
= \;& \sum_{k}^{K}\left(\frac{\sqrt{N}}{K}\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\overline{e}_{ijk\cdot}\right).
\end{aligned}$$

The asymptotic normality can be shown by Lyapounov's Theorem. The Lyapounov condition will be satisfied if

$$L_{AB} = \sum_{k}^{K}\left(\frac{\sqrt{N}}{K}\right)^4 E\left(\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\overline{e}_{ijk\cdot}\right)^4 \to 0.$$

Note that

$$\begin{aligned}
L_{AB} &= \sum_{k}^{K}\left(\frac{\sqrt{N}}{K}\right)^4 E\left(\sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij}\overline{e}_{ijk\cdot}\right)^4 \leq \sum_{k}^{K}\frac{N^2I^3}{K^4}\sum_{i}^{I} E\left(\sum_{j=1}^{J} a_{ij}\overline{e}_{ijk\cdot}\right)^4 \\
&\leq \sum_{k}^{K}\frac{N^2I^3J^3}{K^4}\sum_{i}^{I}\sum_{j=1}^{J} E(a_{ij}^4\overline{e}_{ijk\cdot}^4) = \sum_{k}^{K}\frac{N^2I^3J^3}{K^4}\sum_{i}^{I}\sum_{j=1}^{J}\frac{a_{ij}^4}{n_{ik}^4}E\left(\sum_{l}^{n_{ik}} e_{ijkl}\right)^4 \\
&\leq \sum_{k}^{K}\frac{N^2I^3J^3}{K^4}\sum_{i}^{I}\sum_{j=1}^{J}\frac{a_{ij}^4}{n_{ik}}E(e_{ijkl}^4) = O(K^{-1}),
\end{aligned}$$

where the inequalities follow from Hölder's inequality (2.5.2), and the last equality holds if the fourth moment exists. This completes the proof.

Theorem 3.4.8. *For null hypothesis $H_0(AG)$: all $(\alpha\gamma)_{jk} = 0$ for j=1, ...,J, and k=1, ...,K, let $F_(AG)$ be the statistic given in (3.3.10). If X_{ijkl} has a finite fourth moment, then under $H_0(AG)$,*

$$\frac{\sqrt{K}(F(AG)-1)}{V_{AG}} \xrightarrow{d} N(0,1) \text{ as } K \to \infty,$$

$$V_{AG} = \sqrt{\tau_{AG}}/\sigma_{AG}, \tag{3.4.4}$$

with

$$\begin{aligned}
\tau_{AG} &= \frac{2}{I^2J^2K}\sum_{j,j_1,j_2,j_3}^{J}\sum_{k}^{K}\left[\sum_{i}^{I}\frac{1}{n_{ik}(n_{ik}-1)}\sigma_{i,k,jj_1}\sigma_{i,k,j_2j_3}+\right.\\
&\left.\frac{1}{(I-1)^2}\sum_{i\neq i_1}^{I}\frac{1}{n_{ik}n_{i_1k}}\sigma_{i,k,jj_1}\sigma_{i_1,k,j_2j_3}\right],\\
\sigma_{AG} &= \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{\sigma_{i,k,jj_1}}{n_{ik}}.
\end{aligned}$$

Lemma 3.4.9. *Under the settings and assumptions of Theorem 3.4.8,*

$$MSE_{AG} - \sigma_{AG} \xrightarrow{p} 0 \text{ as } K \to \infty.$$

Proof:

As shown in lemma 3.4.2,

$$E\left[(X_{ijkl} - \overline{X}_{ijk.})(X_{ij_1kl} - \overline{X}_{ij_1k.})\right] = \frac{n_{ik}-1}{n_{ik}}\sigma_{i,k,jj_1}.$$

Then

$$\begin{aligned} & E(MSE_{AG}) \\ = & \frac{1}{IJK}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}E[(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.})] \\ = & \frac{1}{IJK}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{\sigma_{i,k,jj_1}}{n_{ik}} \\ = & \sigma_{AG}. \end{aligned}$$

Therefore,

$$\begin{aligned} & Var(MSE_{AG}) \\ = & \frac{1}{I^2J^2K^2}\sum_{i}^{I}\sum_{k}^{K}\frac{1}{n_{ik}^2(n_{ik}-1)^2}\sum_{l}^{n_{ik}}Var\left[\sum_{j,j_1}^{J}(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.})\right]. \end{aligned}$$

From formula (3.4.2), we have

$$\begin{aligned} & \left|Cov\left[\sum_{j,j_1}^{J}(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.}),\sum_{j_2,j_3}^{J}(X_{ij_2kl}-\overline{X}_{ij_2k.})(X_{ij_3kl}-\overline{X}_{ij_3k.})\right]\right| \\ = & \left|Cov\left[\sum_{j,j_1}^{J}(e_{ijkl}-\overline{e}_{ijk.})(e_{ij_1kl}-\overline{e}_{ij_1k.}),\sum_{j_2,j_3}^{J}(e_{ij_2kl}-\overline{e}_{ij_2k.})(e_{ij_3kl}-\overline{e}_{ij_3k.})\right]\right| \\ < & \infty. \end{aligned}$$

Therefore, we have

$$Var\left[\sum_{j,j_1}^{J}(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.})\right]<\infty.$$

It follows that

$$Var(MSE_{AG})\to 0$$

as $K\to\infty$. Thus we proved $MSE_{AG}-\sigma_{AG}^2\xrightarrow{p}0$ as $K\to\infty$.

Lemma 3.4.10. *Under the settings and assumptions of Theorem 3.4.8 and under $H_0(AG)$, we have*

$$\sqrt{K}(MST_{AG} - P_{AG}(e)) \xrightarrow{p} 0 \text{ as } K \to \infty,$$

where $P_{AG}(e) = \frac{J}{K(I-1)} \sum_{i=1}^{I} \sum_{k=1}^{K} (\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})^2$.

Proof:

Note that under $H_0(AG)$,

$$\begin{aligned}
& MST_{AG} \\
= & \frac{J}{(I-1)(K-1)} \sum_{i=1}^{I} \sum_{k=1}^{K} (\widetilde{e}_{i.k.} - \widetilde{e}_{i...} - \widetilde{e}_{..k.} + \widetilde{e}_{....})^2 \\
= & \frac{J}{(I-1)(K-1)} \sum_{i=1}^{I} \sum_{k=1}^{K} \left[(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})^2 - \frac{2}{K} \sum_{k_1}^{K} (\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.}) + \right. \\
& \left. \frac{1}{K^2} \sum_{k_1}^{K} (\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})^2 \right] \\
= & \frac{J}{(I-1)(K-1)} \sum_{i=1}^{I} \left[\frac{K+1}{K} \sum_{k=1}^{K} (\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})^2 - \frac{2}{K} \sum_{k,k_1}^{K} (\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.}) \right] \\
= & \frac{J}{K(I-1)} \sum_{i=1}^{I} \sum_{k=1}^{K} (\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})^2 - \frac{2J}{K(I-1)(K-1)} \sum_{i=1}^{I} \sum_{k \neq k_1}^{K} (\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.}).
\end{aligned}$$

Thus, we have

$$E[\sqrt{K}(MST_{AG} - P_{AG}(e))] = \frac{2J\sqrt{K}}{K(I-1)(K-1)} \sum_{i=1}^{I} \sum_{k \neq k_1}^{K} E[(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})] = 0.$$

And

$$
\begin{aligned}
& E[\sqrt{K}(MST_{AG} - P_{AG}(e))]^2 \\
= & \frac{4J^2K}{K^2(I-1)^2(K-1)^2} E\left[\sum_{i=1}^{I}\sum_{k\neq k_1}^{K}(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})\right]^2 \\
= & \frac{8J^2}{K(I-1)^2(K-1)^2} E\left[\sum_{(k=k_1)\neq(k_2=k_3)}^{K}\left(\sum_{i=1}^{I}(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_2.} - \widetilde{e}_{..k_2.})\right)\right. \\
& \left.\left(\sum_{i=1}^{I}(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})(\widetilde{e}_{i.k_3.} - \widetilde{e}_{..k_3.})\right)\right] \\
= & \frac{8J^2}{K(I-1)^2(K-1)^2} E\left[\sum_{k\neq k_1}^{K}\sum_{i}^{I}(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})\right) \\
= & \frac{8J^2}{K(I-1)^2(K-1)^2} \sum_{i}^{I}\sum_{k\neq k_1}^{K} E[(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})(\widetilde{e}_{i.k.} - \widetilde{e}_{..k.})]E[(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})(\widetilde{e}_{i.k_1.} - \widetilde{e}_{..k_1.})] \\
= & O(K^{-1}).
\end{aligned}
$$

The last equation holds if X_{ijkl} has the finite fourth moment. Therefore under $H_0(AG)$, $\sqrt{K}(MST_{AG} - P_{AG}(e)) \xrightarrow{p} 0$ as $K \to \infty$.

Proof of Theorem 3.4.8: From Lemma 3.4.9 and Lemma 3.4.10, we need only to consider the asymptotic distribution of $Q_{AG}(e) = \sqrt{K}(P_{AG}(e) - MSE_{AG})$ under $H_0(AG)$.

With some simple algebra, we have

$$
\begin{aligned}
& Q_{AG}(e) \\
= \; & \sqrt{K}\left[\frac{J}{K(I-1)}\sum_{i=1}^{I}\sum_{k=1}^{K}(\widetilde{e}_{i.k.}-\widetilde{e}_{..k.})^2-\frac{1}{IJK}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\right. \\
& \left.\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})(e_{ij_1kl}-\overline{e}_{ij_1k\cdot})\right] \\
= \; & \sqrt{K}\left[\frac{1}{JK(I-1)}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j,j_1}^{J}(\overline{e}_{ijk.}-\widetilde{e}_{.jk.})(\overline{e}_{ij_1k.}-\widetilde{e}_{.j_1k.})-\frac{1}{IJK}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\right. \\
& \left.\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})(e_{ij_1kl}-\overline{e}_{ij_1k\cdot})\right] \\
= \; & \frac{1}{J(I-1)\sqrt{K}}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}(\overline{e}_{ijk.}-\widetilde{e}_{.jk.})(\overline{e}_{ij_1k.}-\widetilde{e}_{.j_1k.})-\frac{1}{IJ\sqrt{K}}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)} \\
& \sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})(e_{ij_1kl}-\overline{e}_{ij_1k\cdot}) \\
= \; & \frac{1}{J(I-1)\sqrt{K}}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}(\overline{e}_{ijk.}\overline{e}_{ij_1k.}-\overline{e}_{ij_1k.}\widetilde{e}_{.jk.}-\overline{e}_{ijk.}\widetilde{e}_{.j_1k.}+\widetilde{e}_{.jk.}\widetilde{e}_{.j_1k.})- \\
& \frac{1}{IJ\sqrt{K}}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(e_{ijkl}e_{ij_1kl}-e_{ij_1kl}\overline{e}_{ijk\cdot}-e_{ijkl}\overline{e}_{ij_1k\cdot}-\overline{e}_{ijk\cdot}\overline{e}_{ij_1k\cdot}) \\
= \; & \frac{1}{J(I-1)\sqrt{K}}\sum_{j,j_1}^{J}\sum_{k}^{K}\left(\sum_{i}^{I}\overline{e}_{ijk.}\overline{e}_{ij_1k.}-\frac{1}{I}\sum_{i,i_1}^{I}\overline{e}_{ijk.}\overline{e}_{i_1j_1k.}\right)- \\
& \frac{1}{IJ\sqrt{K}}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\left(\sum_{l}^{n_{ik}}e_{ijkl}e_{ij_1kl}-\frac{1}{n_{ik}}\sum_{l,l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}\right) \\
= \; & \frac{1}{J(I-1)\sqrt{K}}\sum_{j,j_1}^{J}\sum_{k}^{K}\left(\sum_{i}^{I}\sum_{l,l_1}^{n_{ik}}\frac{I-1}{In_{ik}^2}e_{ijkl}e_{ij_1kl_1}-\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{In_{ik}n_{i_1k}}e_{ijkl}e_{i_1j_1kl_1}\right)- \\
& \frac{1}{IJ\sqrt{K}}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\left(\frac{n_{ik}-1}{n_{ik}}\sum_{l}^{n_{ik}}e_{ijkl}e_{ij_1kl}-\frac{1}{n_{ik}}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}\right) \\
= \; & \frac{1}{IJ\sqrt{K}}\sum_{j,j_1}^{J}\sum_{k}^{K}\left(\sum_{i}^{I}\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}(n_{ik}-1)}e_{ijkl}e_{ij_1kl_1}-\frac{1}{I-1}\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{n_{ik}n_{i_1k}}e_{ijkl}e_{i_1j_1kl_1}\right).
\end{aligned}
$$

Therefore, $E[Q_{AG}(e)] = 0$. It follows that

$$
\begin{aligned}
& Var(Q_{AG}(e)) \\
= \; & \frac{1}{I^2J^2K}\sum_k^K Var\left[\sum_{j,j_1}^J\left(\sum_i^I\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}(n_{ik}-1)}e_{ijkl}e_{ij_1kl_1} - \frac{1}{I-1}\sum_{i\neq i_1}^I\sum_l^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{n_{ik}n_{i_1k}}e_{ijkl}e_{i_1j_1kl_1}\right)\right] \\
= \; & \frac{1}{I^2J^2K}\sum_k^K\left[Var\left(\sum_i^I\sum_{j,j_1}^J\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}(n_{ik}-1)}e_{ijkl}e_{ij_1kl_1}\right) + \right. \\
& \left. Var\left(\sum_{i\neq i_1}^I\sum_{j,j_1}^J\sum_l^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{(I-1)n_{ik}n_{i_1k}}e_{ijkl}e_{i_1j_1kl_1}\right)\right] \\
= \; & \frac{1}{I^2J^2K}\sum_k^K\left[2\sum_i^I\sum_{l\neq l_1}^{n_{ik}}Var\left(\sum_{j,j_1}^J\frac{1}{n_{ik}(n_{ik}-1)}e_{ijkl}e_{ij_1kl_1}\right) + \right. \\
& \left. \frac{2}{(I-1)^2}\sum_{i\neq i_1}^I\left(\sum_l^{n_{ik}}Var\left(\sum_j^J\frac{1}{n_{ik}}e_{ijkl}\right)\right)\left(\sum_{l_1}^{n_{i_1k}}Var\left(\sum_{j_1}^J\frac{1}{n_{i_1k}}e_{i_1j_1kl_1}\right)\right)\right] \\
= \; & \frac{2}{I^2J^2K}\sum_k^K\left[\sum_i^I\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}^2(n_{ik}-1)^2}\sum_{j,j_1,j_2,j_3}^J E(e_{ijkl}e_{ij_2kl})E(e_{ij_1kl_1}e_{ij_3kl_1}) + \right. \\
& \left. \frac{1}{(I-1)^2}\sum_{i\neq i_1}^I\left(\sum_l^{n_{ik}}\frac{1}{n_{ik}^2}\sum_{j,j_1}^J E(e_{ijkl}e_{ij_1kl})\right)\left(\sum_{l_1}^{n_{i_1k}}\frac{1}{n_{i_1k}^2}\sum_{j,j_1}^J E(e_{i_1jkl_1}e_{i_1j_1kl_1})\right)\right] \\
= \; & \frac{2}{I^2J^2K}\sum_k^K\left[\sum_i^I\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j,j_1,j_2,j_3}^J\sigma_{i,k,jj_1}\sigma_{i,k,j_2j_3} + \right. \\
& \left. \frac{1}{(I-1)^2}\sum_{i\neq i_1}^I\left(\frac{1}{n_{ik}}\sum_{j,j_1}^J\sigma_{i,k,jj_1}\right)\left(\frac{1}{n_{i_1k}}\sum_{j,j_1}^J\sigma_{i_1,k,jj_1}\right)\right] \\
= \; & \frac{2}{I^2J^2K}\sum_k^K\left[\sum_i^I\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j,j_1,j_2,j_3}^J\sigma_{i,k,jj_1}\sigma_{i,k,j_2j_3} + \right. \\
& \left. \frac{1}{(I-1)^2}\sum_{i\neq i_1}^I\frac{1}{n_{ik}n_{i_1k}}\sum_{j,j_1,j_2,j_3}^J\sigma_{i,k,jj_1}\sigma_{i_1,k,j_2j_3}\right].
\end{aligned}
$$

Since $Var(Q_{AG}(e))$ is bounded, Lyapunov's condition will be satisfied if

$$
\begin{aligned}
L_{AG}(K) &= \sum_{k}^{K} E \left| \frac{1}{IJ\sqrt{K}} \sum_{j,j_1}^{J} \left(\sum_{i}^{I} \sum_{l \neq l_1}^{n_{ik}} \frac{1}{n_{ik}(n_{ik}-1)} e_{ijkl} e_{ij_1kl_1} - \right. \right. \\
&\qquad \left. \left. \frac{1}{I-1} \sum_{i \neq i_1}^{I} \sum_{l}^{n_{ik}} \sum_{l_1}^{n_{i_1k}} \frac{1}{n_{ik} n_{i_1k}} e_{ijkl} e_{i_1j_1kl_1} \right) \right|^4 \\
&\rightarrow 0.
\end{aligned}
$$

We have

$$
\begin{aligned}
L_{AG}(K) &= \frac{1}{I^4J^4K^2}\sum_{k}^{K} E\left|\sum_{j,j_1}^{J}\left(\sum_{i}^{I}\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}(n_{ik}-1)}e_{ijkl}e_{ij_1kl_1}-\right.\right.\\
&\quad \left.\left.\frac{1}{I-1}\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{n_{ik}n_{i_1k}}e_{ijkl}e_{i_1j_1kl_1}\right)\right|^4\\
&\leq \frac{8J^2}{I^4K^2}\sum_{j,j_1}^{J}\sum_{k}^{K}\left[E\left(\sum_{i}^{I}\sum_{l\neq l_1}^{n_{ik}}\frac{1}{n_{ik}(n_{ik}-1)}e_{ijkl}e_{ij_1kl_1}\right)^4+\right.\\
&\quad \left.E\left(\frac{1}{I-1}\sum_{i\neq i_1}^{I}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}\frac{1}{n_{ik}n_{i_1k}}e_{ijkl}e_{i_1j_1kl_1}\right)^4\right|\\
&\leq \frac{8J^2}{I^4K^2}\sum_{j,j_1}^{J}\sum_{k}^{K}\left[\sum_{i}^{I}\frac{I^3}{n_{ik}^4(n_{ik}-1)^4}E\left(\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}\right)^4+\right.\\
&\quad \left.\sum_{i\neq i_1}^{I}\frac{I^3}{(I-1)n_{ik}^4n_{i_1k}^4}E\left(\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}e_{ijkl}e_{i_1j_1kl_1}\right)^4\right|\\
&\leq \frac{8J^2}{I^4K^2}\sum_{j,j_1}^{J}\sum_{k}^{K}\left[\sum_{i}^{I}\frac{I^3}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}E(e_{ijkl}e_{ij_1kl_1})^4+\right.\\
&\quad \left.\sum_{i\neq i_1}^{I}\frac{I^3}{(I-1)n_{ik}n_{i_1k}}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}E(e_{ijkl}e_{i_1j_1kl_1})^4\right|\\
&= \frac{8J^2}{I^4K^2}\sum_{j,j_1}^{J}\sum_{k}^{K}\left[\sum_{i}^{I}\frac{I^3}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}E(e_{ijkl}^4)E(e_{ij_1kl_1}^4)+\right.\\
&\quad \left.\sum_{i\neq i_1}^{I}\frac{I^3}{(I-1)n_{ik}n_{i_1k}}\sum_{l}^{n_{ik}}\sum_{l_1}^{n_{i_1k}}E(e_{ijkl}^4)E(e_{i_1j_1kl_1}^4)\right|\\
&= O(K^{-1})\ \textit{if the fourth moment of } e_{ijkl} \textit{ exist for any } i,\ j,\ k,\ \textit{and } l.
\end{aligned}
$$

where the two inequalities follow from Hölder's inequality (2.5.2). This completes the proof.

Theorem 3.4.11. *For null hypothesis $H_0(BG)$: all $(\beta\gamma)_{jk} = 0$ for j=1, ...,J, and k=1, ...,K, let $F_(BG)$ be the statistic given in (3.3.12). If X_{ijkl} has a finite fourth moment, then*

under $H_0(BG)$,

$$\frac{\sqrt{K}(F(BG)-1)}{V_{BG}} \xrightarrow{d} N(0,1) \; as K \to \infty, \textit{where } V_{BG} \textit{ is defined in (3.4.5).}$$

The variance component is calculated by

$$V_{BG} = \sqrt{\tau_{BG}}/\sigma_{BG} \tag{3.4.5}$$

where

$$\tau_{BG} = \frac{2}{I^2K(J-1)^2}\sum_{i}^{I}\sum_{k}^{K}\left[\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j,j_1}^{J}\sigma^2_{i,k,jj_1} + \frac{1}{J^2n_{ik}(n_{ik}-1)}\sum_{j,j_1,j_2,j_3}^{J}\sigma_{i,k,jj_1}\sigma_{i,k,j_2j_3} - \right.$$

$$\left.\frac{2}{Jn_{ik}(n_{ik}-1)}\sum_{j,j_1,j_2}^{J}\sigma_{i,k,jj_1}\sigma_{i,k,jj_2}\right]$$

$$\sigma_{BG} = \frac{1}{IK(J-1)}\sum_{i}^{I}\sum_{j}^{J}\sum_{k}^{K}\frac{\sigma^2_{i,k,j}}{n_{ik}} - \frac{1}{IKJ(J-1)}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{\sigma_{i,k,jj_1}}{n_{ik}}.$$

Lemma 3.4.12. *Under the settings and assumptions of Theorem 3.4.11,*

$$MSE_{BG} - \sigma_{BG} \xrightarrow{p} 0 \; as \; K \to \infty.$$

Proof:

As shown in lemma 3.4.2,

$$E\left[(X_{ijkl} - \overline{X}_{ijk.})(X_{ij_1kl} - \overline{X}_{ij_1k.})\right] = \frac{n_{ik}-1}{n_{ik}}\sigma_{i,k,jj_1}.$$

Then

$$\begin{aligned}
& E(MSE_{BG}) \\
= \; & \frac{1}{IK(J-1)}\sum_{i}^{I}\sum_{j}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}E\left[(X_{ijkl} - \overline{X}_{ijk.})^2\right] - \\
& \frac{1}{IKJ(J-1)}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}E\left[(X_{ijkl} - \overline{X}_{ijk.})(X_{ij_1kl} - \overline{X}_{ij_1k.})\right] \\
= \; & \frac{1}{IK(J-1)}\sum_{i}^{I}\sum_{j}^{J}\sum_{k}^{K}\frac{\sigma^2_{i,k,j}}{n_{ik}} - \frac{1}{IKJ(J-1)}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{\sigma_{i,k,jj_1}}{n_{ik}} \\
= \; & \sigma_{BG}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
& Var(MSE_{BG}) \\
= \; & \frac{1}{I^2K^2(J-1)^2}\sum_i^I\sum_k^K\frac{1}{n_{ik}^2(n_{ik}-1)^2}\sum_l^{n_{ik}}\left\{Var\left[\sum_j^J(X_{ijkl}-\overline{X}_{ijk.})^2\right]\right. \\
& +\frac{1}{J^2}Var\left[\sum_{j,j_1}^J(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.})\right] \\
& \left.-\frac{2}{J}Cov\left[\sum_j^J(X_{ijkl}-\overline{X}_{ijk.})^2,\sum_{j,j_1}^J(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.})\right]\right\}.
\end{aligned}
$$

From formula (3.4.2), we have

$$
\begin{aligned}
& \left|Cov\left[\sum_{j,j_1}^J(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.}),\sum_{j_2,j_3}^J(X_{ij_2kl}-\overline{X}_{ij_2k.})(X_{ij_3kl}-\overline{X}_{ij_3k.})\right]\right| \\
= \; & \left|Cov\left[\sum_{j,j_1}^J(e_{ijkl}-\overline{e}_{ijk.})(e_{ij_1kl}-\overline{e}_{ij_1k.}),\sum_{j_2,j_3}^J(e_{ij_2kl}-\overline{e}_{ij_2k.})(e_{ij_3kl}-\overline{e}_{ij_3k.})\right]\right| \\
< \; & \infty.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
& Var\left[\sum_j^J(X_{ijkl}-\overline{X}_{ijk.})^2\right]<\infty \\
& Var\left[\sum_{j,j_1}^J(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.})\right]<\infty \\
& \left|Cov\left[\sum_j^J(X_{ijkl}-\overline{X}_{ijk.})^2,\sum_{j,j_1}^J(X_{ijkl}-\overline{X}_{ijk.})(X_{ij_1kl}-\overline{X}_{ij_1k.})\right]\right|<\infty.
\end{aligned}
$$

Therefore,

$$Var(MSE_{BG})\to 0$$

as $K\to\infty$. It follows that $MSE_{BG}-\sigma_{BG}^2\xrightarrow{p}0$ as $K\to\infty$.

Lemma 3.4.13. *Under the settings and assumptions of Theorem 3.4.11 and under $H_0(BG)$, we have*

$$\sqrt{K}(MST_{BG} - P_{BG}(e)) \xrightarrow{p} 0 \text{ as } K \to \infty,$$

where $P_{BG}(e) = \frac{1}{IK(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (\overline{e}_{ijk.} - \widetilde{e}_{i.k.})^2$.

Proof:

Note that under $H_0(BG)$,

$$\begin{aligned}
& MST_{BG} \\
= & \frac{I}{(J-1)(K-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} \left[(\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})^2 - \frac{2}{K} \sum_{k_1}^{K} (\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.}) + \right. \\
& \left. \frac{1}{K^2} \sum_{k_1}^{K} (\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.})^2 \right] \\
= & \frac{I}{(J-1)(K-1)} \sum_{j=1}^{J} \left[\frac{K+1}{K} \sum_{k=1}^{K} (\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})^2 - \frac{2}{K} \sum_{k,k_1}^{K} (\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.}) \right] \\
= & \frac{I}{K(J-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} (\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})^2 - \frac{2I}{K(J-1)(K-1)} \sum_{j=1}^{J} \sum_{k \neq k_1}^{K} (\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.}).
\end{aligned}$$

And note that

$$\begin{aligned}
& E \left[\frac{I}{K(J-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} (\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})^2 \right] \\
= & \frac{1}{IK(J-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} E \left[\sum_{i=1}^{I} (\overline{e}_{ijk.} - \widetilde{e}_{i.k.}) \right]^2 \\
= & \frac{1}{IK(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} E(\overline{e}_{ijk.} - \widetilde{e}_{i.k.})^2 \\
= & E[P_{BG}(e)].
\end{aligned}$$

Thus, we have

$$E[\sqrt{K}(MST_{BG} - P_{BG}(e))] = \frac{2I\sqrt{K}}{K(J-1)(K-1)} \sum_{j=1}^{J} \sum_{k \neq k_1}^{K} E[(\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.})] = 0.$$

And,

$$
\begin{aligned}
& E[\sqrt{K}(MST_{BG} - P_{BG}(e))]^2 \\
= & \frac{4I^2K}{K^2(J-1)^2(K-1)^2} E\left[\sum_{j=1}^{J}\sum_{k\neq k_1}^{K}(\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.})\right]^2 \\
= & \frac{8I^2}{K(J-1)^2(K-1)^2} E\left[\sum_{(k=k_1)\neq(k_2=k_3)}^{K}\left(\sum_{j=1}^{J}(\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.jk_2.} - \widetilde{e}_{..k_2.})\right)\right. \\
& \left.\left(\sum_{j=1}^{J}(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.})(\widetilde{e}_{.jk_3.} - \widetilde{e}_{..k_3.})\right)\right] \\
= & \frac{8I^2}{K(J-1)^2(K-1)^2} E\left[\sum_{k\neq k_1}^{K}\sum_{j,j_1}^{J}(\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.})(\widetilde{e}_{.j_1k.} - \widetilde{e}_{..k.})(\widetilde{e}_{.j_1k_1.} - \widetilde{e}_{..k_1.})\right) \\
= & \frac{8I^2}{K(J-1)^2(K-1)^2}\sum_{k\neq k_1}^{K}\sum_{j,j_1}^{J} E[(\widetilde{e}_{.jk.} - \widetilde{e}_{..k.})(\widetilde{e}_{.j_1k.} - \widetilde{e}_{..k.})]E[(\widetilde{e}_{.jk_1.} - \widetilde{e}_{..k_1.})(\widetilde{e}_{.j_1k_1.} - \widetilde{e}_{..k_1.})] \\
= & O(K^{-1}).
\end{aligned}
$$

Therefore under $H_0(BG)$, $\sqrt{K}(MST_{BG} - P_{BG}(e)) \xrightarrow{p} 0$ as $K \to \infty$.

Proof of Theorem 3.4.11: From Lemma 3.4.12 and Lemma 3.4.13, we need only to consider the asymptotic distribution of $Q_{BG}(e) = \sqrt{K}(P_{BG}(e) - MSE_{BG})$ under $H_0(BG)$.

With some simple algebra, we have

$$
\begin{aligned}
& Q_{BG}(e) \\
= \; & \sqrt{K}\left[\frac{1}{IK(J-1)}\sum_{i}^{I}\sum_{j}^{J}\sum_{k}^{K}(\overline{e}_{ijk\cdot}-\widetilde{e}_{i\cdot k\cdot})^2-\frac{1}{IK(J-1)}\sum_{i}^{I}\sum_{j}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\right. \\
& \left.\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})^2+\frac{1}{IJK(J-1)}\sum_{i}^{I}\sum_{j,j_1}^{J}\sum_{k}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})(e_{ij_1kl}-\overline{e}_{ij_1k\cdot})\right] \\
= \; & \frac{1}{I\sqrt{K}(J-1)}\sum_{i}^{I}\sum_{k}^{K}\left[\sum_{j}^{J}(\overline{e}_{ijk\cdot}-\widetilde{e}_{i\cdot k\cdot})^2-\sum_{j}^{J}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})^2+\right. \\
& \left.\sum_{j,j_1}^{J}\frac{1}{Jn_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})(e_{ij_1kl}-\overline{e}_{ij_1k\cdot})\right] \\
= \; & \frac{1}{I\sqrt{K}(J-1)}\sum_{i}^{I}\sum_{k}^{K}\left[(\sum_{j}^{J}\overline{e}_{ijk\cdot}^2-\frac{1}{J}\sum_{j,j_1}^{J}\overline{e}_{ijk\cdot}\overline{e}_{ij_1k\cdot})-\sum_{j}^{J}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})^2+\right. \\
& \left.\sum_{j,j_1}^{J}\frac{1}{Jn_{ik}(n_{ik}-1)}\sum_{l}^{n_{ik}}(e_{ijkl}-\overline{e}_{ijk\cdot})(e_{ij_1kl}-\overline{e}_{ij_1k\cdot})\right] \\
= \; & \frac{1}{I\sqrt{K}(J-1)}\sum_{i}^{I}\sum_{k}^{K}\left[(\frac{1}{n_{ik}^2}\sum_{j}^{J}\sum_{l,l_1}^{n_{ik}}e_{ijkl}e_{ijkl_1}-\frac{1}{Jn_{ik}^2}\sum_{j,j_1}^{J}\sum_{l,l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1})-\right. \\
& (\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j}^{J}\sum_{l}^{n_{ik}}e_{ijkl}^2-\frac{1}{n_{ik}^2(n_{ik}-1)}\sum_{j}^{J}\sum_{l,l_1}^{n_{ik}}e_{ijkl}e_{ijkl_1})+ \\
& \left.(\frac{1}{Jn_{ik}(n_{ik}-1)}\sum_{j,j_1}^{J}\sum_{l}^{n_{ik}}e_{ijkl}e_{ij_1kl}-\frac{1}{Jn_{ik}^2(n_{ik}-1)}\sum_{j,j_1}^{J}\sum_{l,l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1})\right] \\
= \; & \frac{1}{I\sqrt{K}(J-1)}\sum_{i}^{I}\sum_{k}^{K}\left[\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ijkl_1}-\frac{1}{Jn_{ik}(n_{ik}-1)}\sum_{j,j_1}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}\right].
\end{aligned}
$$

Therefore, $E[Q_{BG}(e)] = 0$. It follows that

$$
\begin{aligned}
& Var(Q_{BG}(e)) \\
= & \frac{1}{I^2K(J-1)^2}\sum_{i}^{I}\sum_{k}^{K} Var\left[\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j}^{J}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ijkl_1} - \frac{1}{Jn_{ik}(n_{ik}-1)}\sum_{j,j_1}^{J}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ij_1kl_1}\right] \\
= & \frac{2}{I^2K(J-1)^2}\sum_{i}^{I}\sum_{k}^{K}\sum_{l\neq l_1}^{n_{ik}}\left[Var(\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j}^{J} e_{ijkl}e_{ijkl_1}) + Var(\frac{1}{Jn_{ik}(n_{ik}-1)}\right. \\
& \left.\sum_{j,j_1}^{J} e_{ijkl}e_{ij_1kl_1}) - 2Cov(\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j}^{J} e_{ijkl}e_{ijkl_1}, \frac{1}{Jn_{ik}(n_{ik}-1)}\sum_{j,j_1}^{J} e_{ijkl}e_{ij_1kl_1})\right] \\
= & \frac{2}{I^2K(J-1)^2}\sum_{i}^{I}\sum_{k}^{K}\sum_{l\neq l_1}^{n_{ik}}\left[\frac{1}{n_{ik}^2(n_{ik}-1)^2}\sum_{j,j_1}^{J} E(e_{ijkl}e_{ij_1kl})E(e_{ijkl_1}e_{ij_1kl_1}) + \right. \\
& \frac{1}{J^2n_{ik}^2(n_{ik}-1)^2}\sum_{j,j_1,j_2,j_3}^{J} E(e_{ijkl}e_{ij_2kl})E(e_{ij_1kl_1}e_{ij_3kl_1}) - \\
& \left.\frac{2}{J^2n_{ik}^2(n_{ik}-1)^2}\sum_{j,j_1,j_2}^{J} E(e_{ijkl}e_{ijkl_1}e_{ij_1kl}e_{ij_2kl_1})\right] \\
= & \frac{2}{I^2K(J-1)^2}\sum_{i}^{I}\sum_{k}^{K}\left[\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j,j_1}^{J}\sigma_{i,k,jj_1}^2 + \frac{1}{J^2n_{ik}(n_{ik}-1)}\sum_{j,j_1,j_2,j_3}^{J}\sigma_{i,k,jj_1}\sigma_{i,k,j_2j_3} - \right. \\
& \left.\frac{2}{Jn_{ik}(n_{ik}-1)}\sum_{j,j_1,j_2}^{J}\sigma_{i,k,jj_1}\sigma_{i,k,jj_2}\right].
\end{aligned}
$$

Since $Var(Q_{BG}(e))$ is bounded, Lyapunov's condition will be satisfied if

$$
\begin{aligned}
L_{BG}(K) &= \sum_{k=1}^{K} E\left|\frac{1}{I\sqrt{K}(J-1)}\sum_{i}^{I}[\frac{1}{n_{ik}(n_{ik}-1)}\sum_{j}^{J}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ijkl_1} - \right. \\
&\quad \left.\frac{1}{Jn_{ik}(n_{ik}-1)}\sum_{j,j_1}^{J}\sum_{l\neq l_1}^{n_{ik}} e_{ijkl}e_{ij_1kl_1}]\right|^4 \\
&\to 0.
\end{aligned}
$$

We have

$$
\begin{aligned}
L_{BG}(K) &= \frac{1}{I^4K^2(J-1)^4}\sum_{k=1}^{K}E\left|\sum_{i}^{I}\frac{1}{n_{ik}(n_{ik}-1)}[\sum_{j}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ijkl_1}-\frac{1}{J}\sum_{j,j_1}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}]\right|^4 \\
&\leq \frac{1}{IK^2(J-1)^4}\sum_{i}^{I}\sum_{k=1}^{K}\frac{1}{n_{ik}^4(n_{ik}-1)^4}E\left|\sum_{j}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ijkl_1}-\frac{1}{J}\sum_{j,j_1}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1}\right|^4 \\
&\leq \frac{8}{IK^2(J-1)^4}\sum_{i}^{I}\sum_{k=1}^{K}\frac{1}{n_{ik}^4(n_{ik}-1)^4}\left[E(\sum_{j}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ijkl_1})^4+E(\frac{1}{J}\sum_{j,j_1}^{J}\sum_{l\neq l_1}^{n_{ik}}e_{ijkl}e_{ij_1kl_1})^4\right] \\
&\leq \frac{8}{IK^2(J-1)^4}\sum_{i}^{I}\sum_{k=1}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}\left[E(\sum_{j}^{J}e_{ijkl}e_{ijkl_1})^4+\frac{1}{J^4}E(\sum_{j,j_1}^{J}e_{ijkl}e_{ij_1kl_1})^4\right] \\
&\leq \frac{8}{IK^2(J-1)^4}\sum_{i}^{I}\sum_{k=1}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}\left[J^3\sum_{j}^{J}E(e_{ijkl}e_{ijkl_1})^4+J^2\sum_{j,j_1}^{J}E(e_{ijkl}e_{ij_1kl_1})^4\right] \\
&= \frac{8}{IK^2(J-1)^4}\sum_{i}^{I}\sum_{k=1}^{K}\frac{1}{n_{ik}(n_{ik}-1)}\sum_{l\neq l_1}^{n_{ik}}\left[J^3\sum_{j}^{J}E(e_{ijkl}^4)E(e_{ijkl_1}^4)+\right. \\
&\quad \left.J^2\sum_{j,j_1}^{J}E(e_{ijkl}^4)E(e_{ij_1kl_1}^4)\right] \\
&= O(K^{-1})\ if\ the\ fourth\ moment\ of\ e_{ijkl}\ exist\ for\ any\ i,\ j,\ k,\ and\ l,
\end{aligned}
$$

where the two inequalities follow from Hölder's inequality (2.5.2). This completes the proof.

3.5 Simulation results

Simulation study was carried out to evaluate the proposed nonparametric test statistics (NPT) in various conditions. First, we looked at their type I error rates under null hypothesis. To test the robustness of the proposed statistics, we generated random numbers from various distributions and covariance structures. Secondly, power analysis were conducted to compare NPT statistics to linear mixed-effects model (LME) and generalized estimating equations (GEE). In order to generate random numbers as close as to real microarray data, we use bootstrap to re-sample data from a two-treatment microarray experiment.

Proper within-subject correlation was then incorporated to the data, and power curves were produced for each of NPT, LME, and GEE methods.

All the data in this section were generated from the model specified in (3.2.1)

$$X_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijkl}.$$

Throughout the manuscript, all calculations and simulations were completed with R programming. LME and GEE methods were implemented by using *gls* and *geese* functions from R packages *nlme* and *geepack*, respectively ((Pinheiro and Bates (2000)); (Yan and Fine (2004)).

3.5.1 Type I error rate analysis with simulated data

In this section, we measured the specificity of our proposed model (NPT) based on type I error rates for simulated data from various distributions. In most of microarray applications, the researchers are interested in identifying gene expression changed by treatment comparing to a control group. In the simulation, we will focused on two treatment groups. The number of time points we simulated is either 2 or 5. As balanced design is only a special form of unbalanced design, here we only consider unbalanced design in that four fifths of genes having 4 replications and the remaining one fifth of genes having 6 replications.

The gene expression microarray data were often modeled with log-normal or normal distribution (Sidorov et al. (2002); Hoyle et al. (2002)). Nonetheless, there are a number of arguments that the real gene profile does not closely fit these distributions, and to model the distribution is still an hot research field with big challenge (Kerr et al. (2000); Konishi (2004)). To allow a wide variety of data types, normal, exponential, Poisson, and Cauchy distributions were used to generate random samples. Appropriate within-subject correlation structures were introduced into the data with the methods described for each experiment. Throughout the manuscript, all simulations were performed using 1000 iterations.

For normal, exponential, and Poisson distributions, the mean of random numbers was set to 2. The normal distribution was given a standard deviation of 1. The Cauchy distribution had a location value of 0, and a scale value of 1. The within-subject correlation (over time points) were modeled either with AR(1) or unstructured correlation structure. For AR(1) correlation, the covariance vector X was conditioned with $cov(X_{ijk}, X_{ij_1k}) = .5^{|j-j_1|}$. The unstructured correlation structures were obtained by generating correlation symmetric matrix that has random numbers uniformly distributed between 0 and 1.

First, we examined the proposed test statistic for $H_0(A)$ of no gene expression variations across treatments. A data matrix X of J rows and n columns were were randomly generated with each row representing a time point and each column representing a gene. J is the number of time points, and n is the sum of the number of replications for all genes across all treatment groups. The data matrix were generated under null hypothesis such that there were no expression changes across columns. An AR(1) or unstructured correlation structure matrix L with J rows and J columns were then generated as described above. We used the Cholesky decomposition to produce the lower triangle half matrix h for the correlation matrix L. The Cholesky decomposition is conducted via R function *chol*. Thus the data matrix Y=h*X has the desired correlation structure and it would be used for subsequent data analysis. The matrix Y had equal means across columns. Nonetheless, at different time points (across rows), the values from the same gene could be varied. Random data generation from Poisson distribution was performed slightly differently. We aimed to use Poisson distribution to generate integer data that simulate specific image data, such as DNA copy number and cell count. Therefore, we first calculated the means matrix M of the same dimensions as Y. The matrix M was generated by incorporating the correlation structure to a matrix with identical elements of value 2. Then the random data matrix Y was obtained by generate random numbers from Poisson distribution with mean M.

The results of Type I error rate at alpha level 0.05 were given in Table 3.1. Subjects were assumed to be repeatedly measured at 5 time points with AR(1) within-subject correlation

structure. As the number of genes increases, the error rate converges to 0.05 for normal, exponential, and Poisson distribution. They were close to 0.05 with at least 30 genes. However, the error rate for Cauchy distribution did never go to 0.05 because it does not have a finite mean.

no.gene	normal	exponential	Poisson	Cauchy
5	0.051	0.071	0.075	0.035
10	0.056	0.067	0.062	0.026
20	0.037	0.058	0.053	0.023
30	0.049	0.053	0.055	0.023
40	0.053	0.057	0.054	0.020
50	0.057	0.053	0.052	0.026
100	0.041	0.055	0.051	0.025
200	0.049	0.056	0.054	0.020
500	0.051	0.049	0.047	0.021
1000	0.051	0.051	0.051	0.020

Table 3.1: Estimated type I error estimate of the test of no treatment effect at 0.05 level in unbalanced design. The data from the same gene follows AR(1) with correlation =0.5.

Table 3.2 showed the type I error rates for unstructured correlation. Either 2 or 5 time points were simulated for each dataset. For normal, exponential, and Poisson distributions, the error rates for at least 5 genes and 2 or 5 time points were in high agreement with the expected level $alpha = 0.05$. The error rate for Cauchy distribution failed to converge to 0.05.

Secondly, we conducted hypothesis test for the time effect with simulated data. The random numbers were generated as described above. As correlation introduction via the Cholesky decomposition does not maintain equal means across time points (rows), we used an iterative algorithm to generate the AR(1) or unstructured correlation. Suppose for gene i, the correlation between the jth and (j+1)th time points is ρ that was given based on AR(1) or unstructured correlation. Given the value X_j of gene i at the jth time point, the

no.time	no.gene	normal	exponential	Poisson	Cauchy
2	5	0.060	0.053	0.063	0.021
	10	0.047	0.052	0.048	0.024
	20	0.053	0.046	0.054	0.026
	30	0.048	0.063	0.058	0.019
	40	0.044	0.052	0.053	0.021
	50	0.043	0.052	0.057	0.020
	100	0.040	0.050	0.042	0.020
5	5	0.056	0.052	0.059	0.032
	10	0.053	0.055	0.057	0.025
	20	0.047	0.045	0.066	0.020
	30	0.060	0.058	0.050	0.014
	40	0.050	0.049	0.047	0.018
	50	0.044	0.041	0.041	0.016
	100	0.062	0.047	0.050	0.023

Table 3.2: Estimated type I error estimate of the test of no treatment effect at 0.05 level in unbalanced design. The data from the same gene have unstructured correlation.

random copy number of the (j+1)th time point can be obtained by

$$X_{j+1} = \rho X_j + b,$$

where b is a random number with the mean of $2(1-\rho)$. Thus the mean of X_{j+1} is 2, which is the same as that of X_j. For Poisson distribution, we first generated the mean values with the iterative algorithm, and then used the means to generate random integer numbers.

The type I error rates at alpha level 0.05 with unstructured correlation structure were shown in Table 3.3. Two time points were simulated for each experiment. The error rates were close to 0.05 for normal, exponential, and Poisson distributions when there were at least 50 genes in the dataset. The error rate for Cauchy distribution did not converge to 0.05.

Thirdly, simulation was conducted to test the gene effect. Under null hypothesis of no gene log-ratio variation, the data generating process was the same to that for test of the treatment effect. An unstructured correlation was introduced to the repeated measures for

no.gene	normal	exponential	Poisson	Cauchy
5	0.049	0.058	0.061	0.025
10	0.062	0.050	0.041	0.016
20	0.055	0.055	0.045	0.018
30	0.074	0.057	0.058	0.020
40	0.062	0.040	0.051	0.018
50	0.047	0.052	0.054	0.022
100	0.042	0.040	0.044	0.017
200	0.044	0.046	0.045	0.021
500	0.053	0.043	0.041	0.017
1000	0.056	0.046	0.037	0.017

Table 3.3: Estimated type I error of the test of no time effect at 0.05 level. The data from the same gene followed unstructured correlation. For each simulation, there are 2 time points.

either 2 or 5 time points for each gene. The type I error rates at alpha level 0.05 were shown in Table 3.4. Normal, exponential, and Poisson distributions with both 2 and 5 time points had similar convergence rate. They converged to 0.05 with at least 40 genes. Cauchy distribution did not converge to 0.05 as expected.

The next test was concerned with the interaction of treatment and time effect. Under null hypothesis of no interaction, we generate random data with the same process as to that for test of the time effect. An unstructured correlation was introduced to the repeated measures of two time points for each gene. The type I error rates at alpha level 0.05 were shown in Table 3.5. Normal, exponential, and Poisson distributions had error rates close to 0.05 when the number of genes was above 50. Cauchy distribution did not converge to 0.05.

The fifth test was conducted for the interaction of treatment and the gene effect. Under null hypothesis of no interaction, the data were simulated in the same way as to that for test of the gene effect. An unstructured correlation was introduced to the repeated measures of two time points for each gene. The type I error rates at alpha level 0.05 were shown in Table 3.6. Normal, exponential, and Poisson distributions had error rates converging to 0.05 with at least 50 genes. Cauchy distribution did not converge to 0.05.

no.time	no.gene	normal	exponential	Poisson	Cauchy
2	5	0.076	0.093	0.097	0.184
	10	0.063	0.067	0.058	0.145
	20	0.062	0.069	0.055	0.141
	30	0.055	0.063	0.053	0.177
	40	0.067	0.055	0.054	0.138
	50	0.059	0.057	0.053	0.135
	100	0.052	0.040	0.049	0.158
	200	0.049	0.052	0.044	0.141
	500	0.060	0.058	0.052	0.137
	1000	0.037	0.055	0.048	0.141
5	5	0.073	0.096	0.072	0.189
	10	0.080	0.068	0.074	0.179
	20	0.062	0.062	0.065	0.135
	30	0.067	0.064	0.070	0.148
	40	0.051	0.060	0.046	0.140
	50	0.043	0.050	0.047	0.159
	100	0.053	0.056	0.038	0.143
	200	0.046	0.046	0.047	0.146
	500	0.041	0.057	0.041	0.113
	1000	0.060	0.063	0.027	0.142

Table 3.4: Estimated type I error rates of the test of no gene effect at 0.05 level. The data from the same gene follow unstructured correlation. There were either 2 or 5 time points for repeated measures.

Our last attempt was for the interaction test of time and gene effect. Under null hypothesis of no interaction, the random samples were generated in the same way as described for test of time effect. An unstructured correlation was introduced to the repeated measures for either 2 or 5 time points for each gene. The number of time points did not play an important role to affect the error rate. The type I error rates at alpha level 0.05 were shown in Table 3.7. Normal, exponential, and Poisson distributions had error rates converging to 0.05 with at least 40 genes. Cauchy distribution did not converge to 0.05.

no.gene	normal	exponential	Poisson	Cauchy
5	0.087	0.103	0.099	0.046
10	0.074	0.082	0.064	0.035
20	0.061	0.063	0.050	0.024
30	0.070	0.071	0.063	0.019
40	0.071	0.060	0.065	0.019
50	0.064	0.052	0.056	0.011
100	0.037	0.051	0.048	0.012
200	0.043	0.050	0.052	0.018
500	0.048	0.040	0.051	0.022
1000	0.057	0.046	0.048	0.013

Table 3.5: Estimated type I error of the test of no treatment*time interaction at 0.05 level. The data from the same gene followed unstructured correlation. For each simulation, there are 2 time points.

no.gene	normal	exponential	Poisson	Cauchy
5	0.088	0.083	0.089	0.195
10	0.072	0.076	0.064	0.173
20	0.070	0.072	0.059	0.186
30	0.053	0.063	0.049	0.167
40	0.053	0.066	0.050	0.138
50	0.058	0.059	0.056	0.163
100	0.057	0.064	0.043	0.140
200	0.054	0.053	0.049	0.139
500	0.055	0.050	0.052	0.147
1000	0.044	0.049	0.051	0.139

Table 3.6: Estimated type I error of the test of no treatment*gene interaction at 0.05 level. The data from the same gene followed unstructured correlation. For each simulation, there are 2 time points.

3.5.2 Power analysis with bootstrap data and simulated data

In this sub-section, power analysis was conducted to compare the proposed method (NPT) with linear mixed model (LME) and generalized estimating equations (GEE) . Since the distribution fitting for microarray data is not satisfactory as discussed in section 3.1, we primarily used bootstrap to produce data samples based on real microarray data. As the

no.time	no.gene	normal	exponential	Poisson	Cauchy
2	5	0.083	0.099	0.107	0.261
	10	0.055	0.068	0.090	0.254
	20	0.056	0.053	0.068	0.248
	30	0.051	0.069	0.073	0.242
	40	0.047	0.050	0.050	0.238
	50	0.049	0.039	0.068	0.251
	100	0.042	0.037	0.057	0.242
	200	0.032	0.034	0.037	0.270
	500	0.031	0.033	0.039	0.254
	1000	0.034	0.034	0.028	0.249
5	5	0.090	0.088	0.097	0.261
	10	0.049	0.072	0.074	0.267
	20	0.047	0.043	0.049	0.252
	30	0.053	0.050	0.039	0.242
	40	0.045	0.054	0.034	0.246
	50	0.034	0.037	0.039	0.245
	100	0.039	0.037	0.048	0.258
	200	0.055	0.045	0.032	0.259
	500	0.038	0.038	0.029	0.249
	1000	0.028	0.037	0.041	0.239

Table 3.7: Estimated type I error rates of the test of no time*gene effect at 0.05 level. The data from the same gene follow unstructured correlation. There were either 2 or 5 time points for repeated measures.

simulation study for type I error rates, we only used unstructured correlation matrix for it is a general form. The correlation was then introduced to bootstrap samples with the iterative algorithm described in the previous sub-section. In each experiment, we only consider 2 time points for it is common in a longitudinal microarray experiment design. Based on our simulation study for type I error rates, 50 genes is sufficient large to achieve expected error rates. So we used 50 genes for all power analysis experiments. Similar to type I error simulation, the design was unbalanced with four fifths of genes having 4 replications and one fifth of genes having 6 replications. Two treatment groups were considered in this bootstrap study. As a comparison, we also used random number generated from normal distributions for power analysis. The results were shown at the end of the section for two hypothesis

tests.

We acquired the data of two microarray samples for IL-2 response experiment in murine T cell (Zhang et al. (2007)). One sample were stimulated with IL-2 for 4 hours, and the other was a control without IL-2 treatment. The detail of the experiment design and data preprocession was described in the next section (3.6). We used the following procedure to determine the gene list under H_0 and under H_a. For each gene, we calculated its log-ratio expressions of the IL-2 treatment to the control. If the log-ratio was between ± 0.1, we regarded the gene is normal (under H_0). If the log-ratio was above 1.3, we regarded the gene to be abnormal (under H_a) for it was activated by IL-2 by more than 3.5 fold change. In such an arbitrary definition, we had 3652 genes in the normal list, and 1409 genes in the abnormal list. The power analysis was conducted by contaminating a bootstrap sample of normal genes with a small proportion of a bootstrap sample of abnormal genes. For different tests, we bootstrapped either the original gene expressions (log transformed) or the log ratios of the two samples. We will discuss which type of data should be used for each experiment.

LME analysis was carried out by *gls* function available in contributed R package *nlme* (Pinheiro and Bates (2000)). Unstructured within-subject correlation structure was assumed to the model. Since random effect was not considered, we used generalized least squares were used to fit the LME model (Carroll and Ruppert (1988)). GEE analysis was carried out by *geese* function available in contributed R package *geepack* (Yan and Fine (2004)). Unstructured correlation was assumed to the model. The Gaussian family was assumed for the error distribution.

First, we conducted power analysis for the treatment effect. We considered log ratio data in this test to simulate the data of two-channel microarray (two dyes) or the data comparing to a reference sample. In the simulation, for one treatment group, all data came from the normal gene lists. For the other treatment group, the normal genes were contaminated a

small percentage of abnormal genes. Both normal and abnormal data were bootstrapped from the previously defined gene lists. The contamination percentage in each dataset varied from 0 to 1%. The unstructured correlation structure was introduced to the bootstrap data by the Cholesky decomposition as described in the previous sub-section. The three power curves for NPT, LME, and GEE were shown in Figure 3.1. When the contamination was less than 0.7%, GEE had a higher power than the other two methods, but its power did not monotonously increase with percentage of contaminations. NPT outperformed LME and GEE when there was at least 0.7% contamination, and it was the method that first reached 80% power. Therefore, NPT had a better performance in this situation.

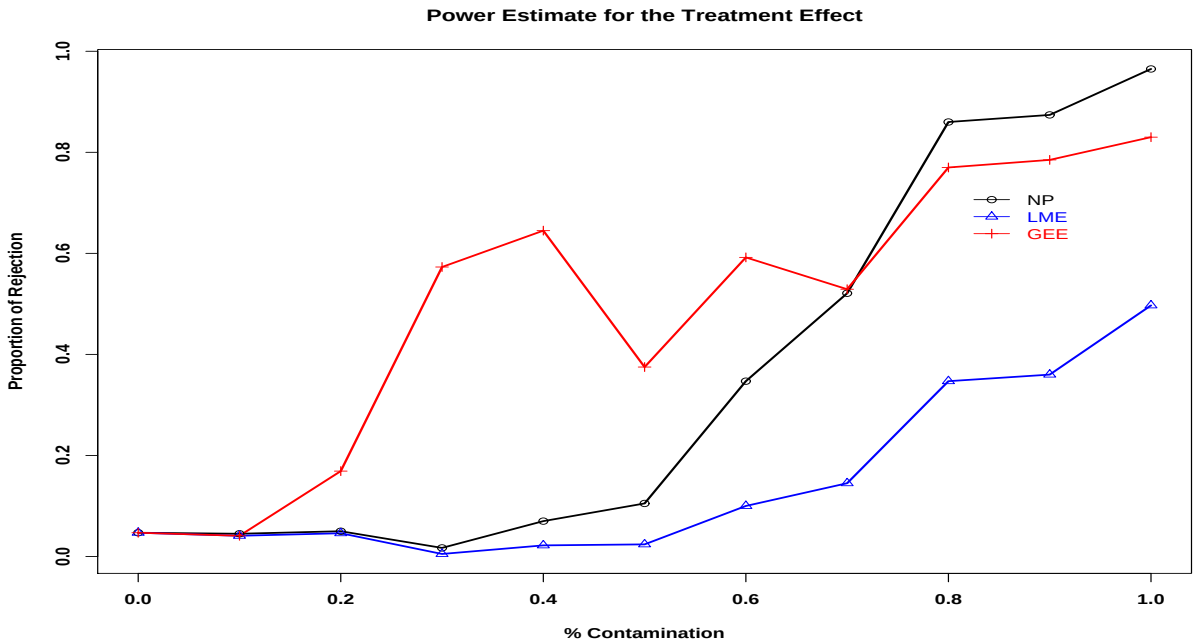


Figure 3.1: *The power curves of testing the treatment effect for unbalanced design with unstructured correlation correlation. There were 2 time points in the experiment.*

To illustrate how the number of genes affects statistical power, we displayed power curves with varying number of genes. The contamination percentage was fixed at 0.5%. As the number of genes increased, the power of NPT approached to 100% (Figure 3.2). It was the only method whose power significantly increased with the number of genes. GEE increased slightly. The power of LME stayed close to zero for the whole range of number of genes.

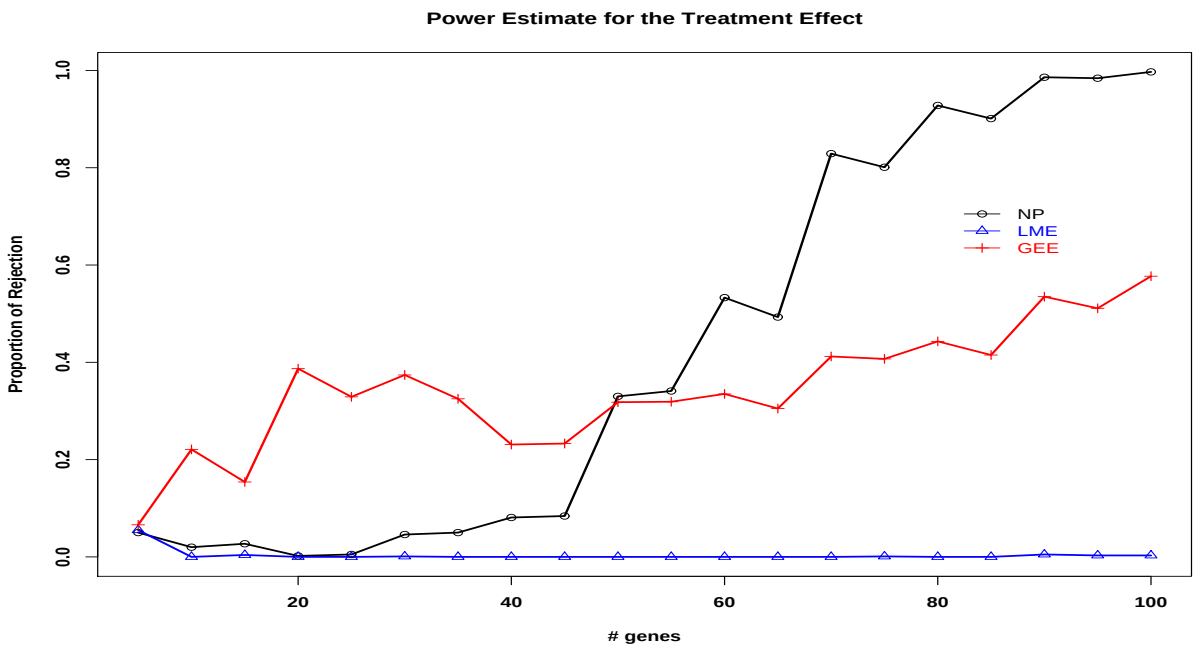


Figure 3.2: *The power curves of testing the treatment effect with 0.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

Our next test was for time effect. We used the log expression values for bootstrap. Since we only considered two time points, we bootstrapped the gene IDs, and let the data from IL-2 sample form one time point, and let the data from the control sample form the other time point. As discussed in the previous subsection, we had to use the iterative algorithm to incorporate the unstructured correlation. All three methods were very sensitive in detecting gene expression variation over time. At 2% contamination, all reached a power of 100% (Figure 3.3). Nonetheless, NPT performed the best because it was the first method to reach 100% power.

Thirdly, we explored the sensitivity of testing the gene effect with the three methods. As described in section 3.3, log ratio is commonly used to identify differentially expressed genes. So we calculated the log ratio for this hypothesis test. Both treatment groups were contaminated with the same percentage ($\leq 1\%$) of abnormal genes. From Figure 3.4, NPT showed much higher power than LME and GEE with minimal contamination. With 0.5%

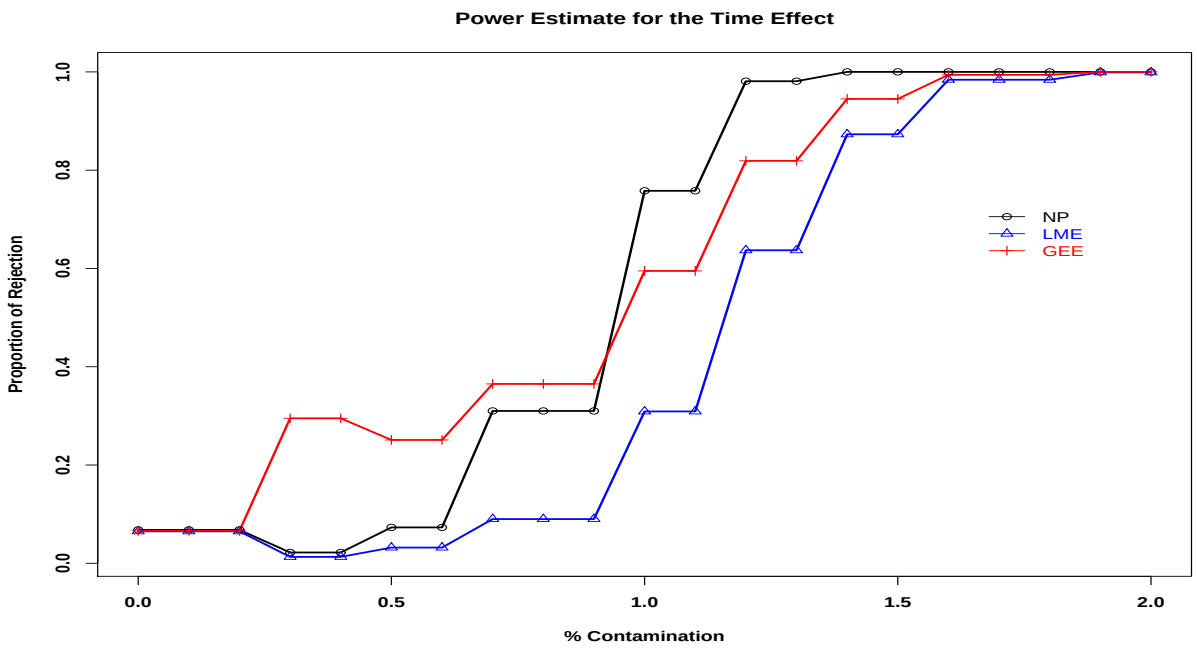


Figure 3.3: *The power curves of testing time effect with up to 2% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

contamination, NPT has 89.2% power, whereas LME and GEE have powers of 31.4% and 24.3%, respectively.

Fourthly, we conducted power simulation analysis for treatment and time interaction. The data generation is similar to that for time effect except that only one treatment group was contaminated with abnormal genes in this study. We ran simulation up to 1.5% contamination to illustrate how the three power curves approached to 100%. As shown in Figure 3.5, the power curves of the three methods were very close thought NPT reached 80% and 100% powers first.

The fifth test was for treatment and gene interaction. Bootstrap data were generated in a similar way to that for test of the gene effect. The only exception is that we only contaminated one treatment group with up to 2% of abnormal genes. The performance of NPT was obviously higher than the other two methods (Figure 3.6). The power of NPT went to above 90% at 1% contamination, whereas the powers of GEE and LME reached

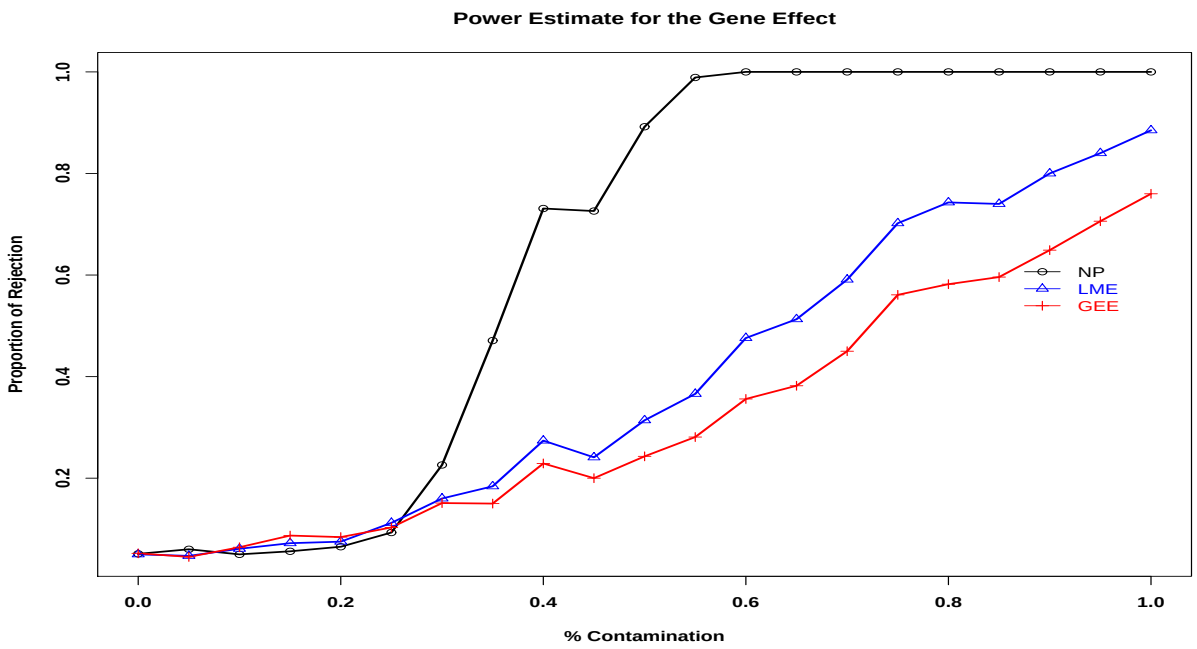


Figure 3.4: *The power curves of testing time effect with up to 1% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

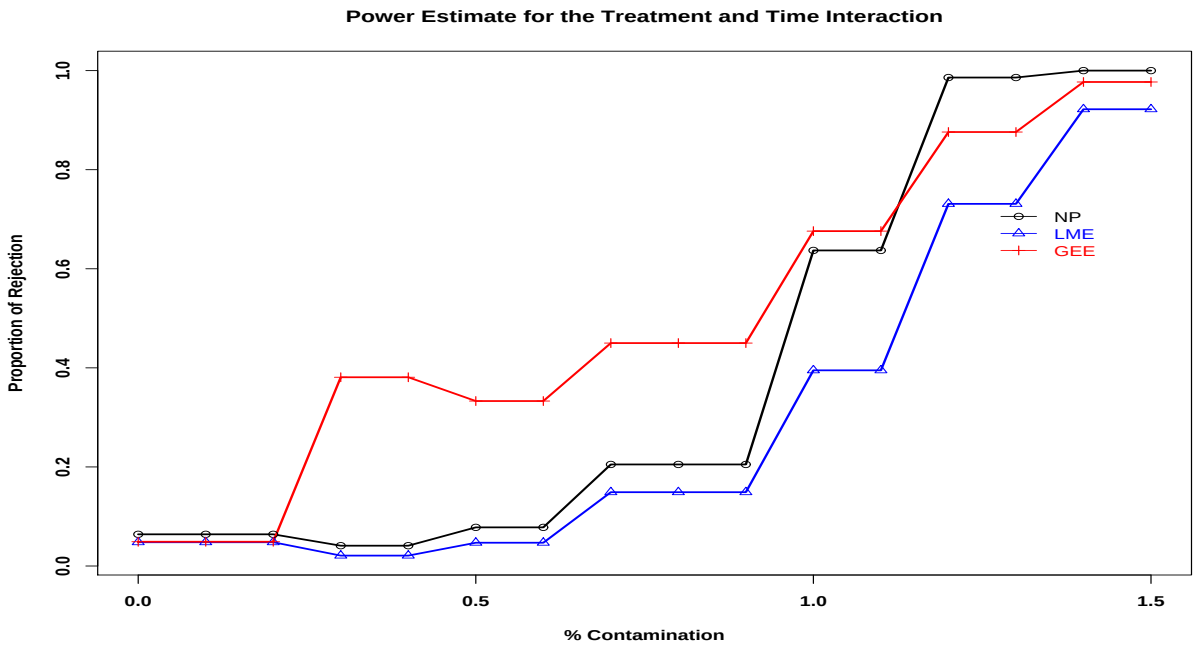


Figure 3.5: *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

90% at close to 2% contamination.

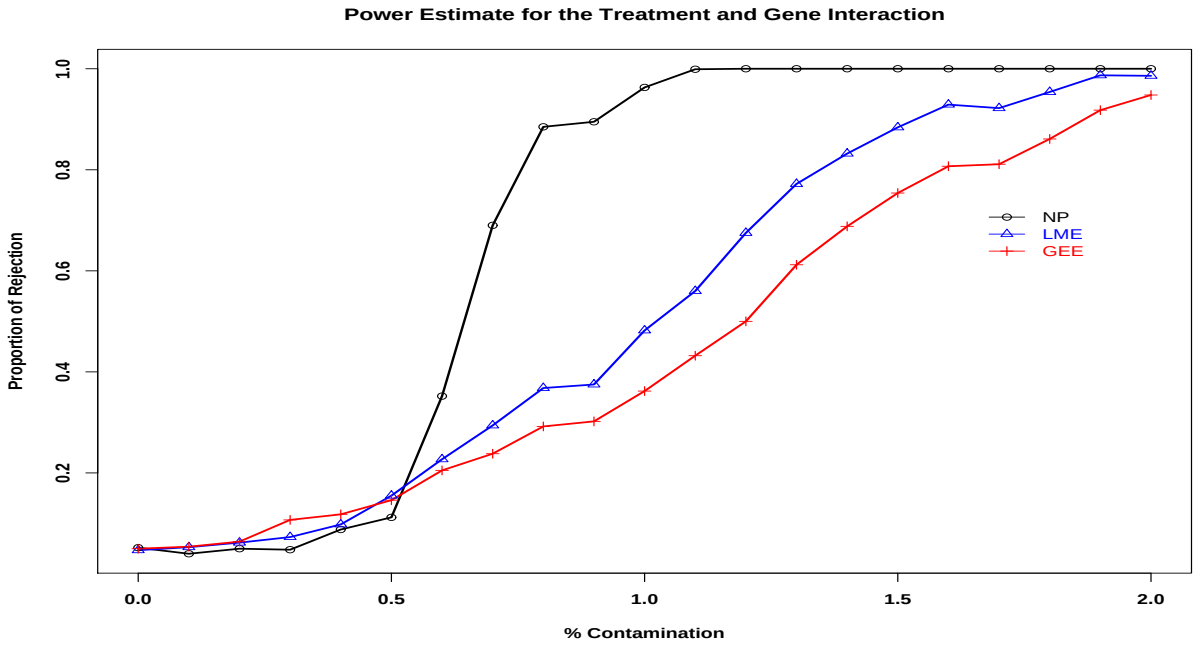


Figure 3.6: *The power curves of testing time effect with up to 2% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

Our sixth test was conducted for time and gene interaction. We used the same data-generating process as for test of time effect. We ran the simulation up to 1.5% contamination with two time points. As the other bootstrap studies, NPT performed the best in the power analysis (Figure 3.6). Its power curve increased sharply above 90% at about 0.5% contamination. GEE and LME achieved above 90% power with at least 1% contamination.

To evaluate the effects of real microarray data on the statistical powers of these methods, we generated Gaussian distributed random numbers for power analysis. We want to know the behaviors of the three methods in such an ideal condition.

The simulations were only conducted for the time effect and the time*gene interaction. For the test of the time effect, the data for one time point were generated from a normal distribution with mean 0 and standard deviation 1. The data for the other time point were based on a normal distribution with a mean varying from 0 to 0.4. Its standard deviation

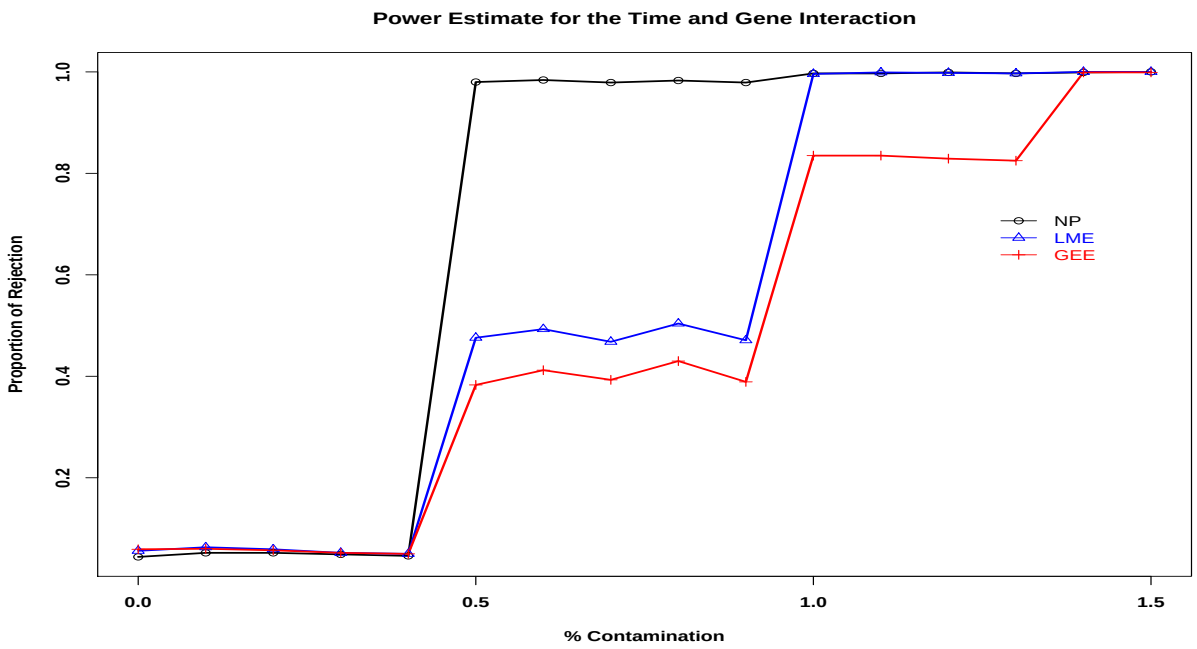


Figure 3.7: *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

was still 1. For the test of time*gene interaction, the data for the majority of genes came from a normal distribution with mean 0 and standard deviation 1. The remaining gene expressions had a mean of 2 and a standard deviation of 1. As shown in Figures 3.8 and 3.9, the three methods behaved equivalently well for the normality assumption.

Of all the bootstrap simulations, NPT were constantly the first method to reach a high power ($> 90\%$). In most of the conditions, especially for those tests for the gene effect or its interaction, it performed significantly better than LME and GEE. As shown in Figure 3.2, whereas the large number of genes may have a negative effect on the performance of LME and GEE, we expected the performance of NPT increases with the number of genes. NPT is a robust method as well. It performed as good as LME and GEE for Gaussian distributed data. For the noisy intensity data, it maintains a high sensitivity to detect a very low contamination (usually $< 1\%$ contamination) with above 80% power. As we used real microarray data for bootstrap, we expect NPT have similar high performance in

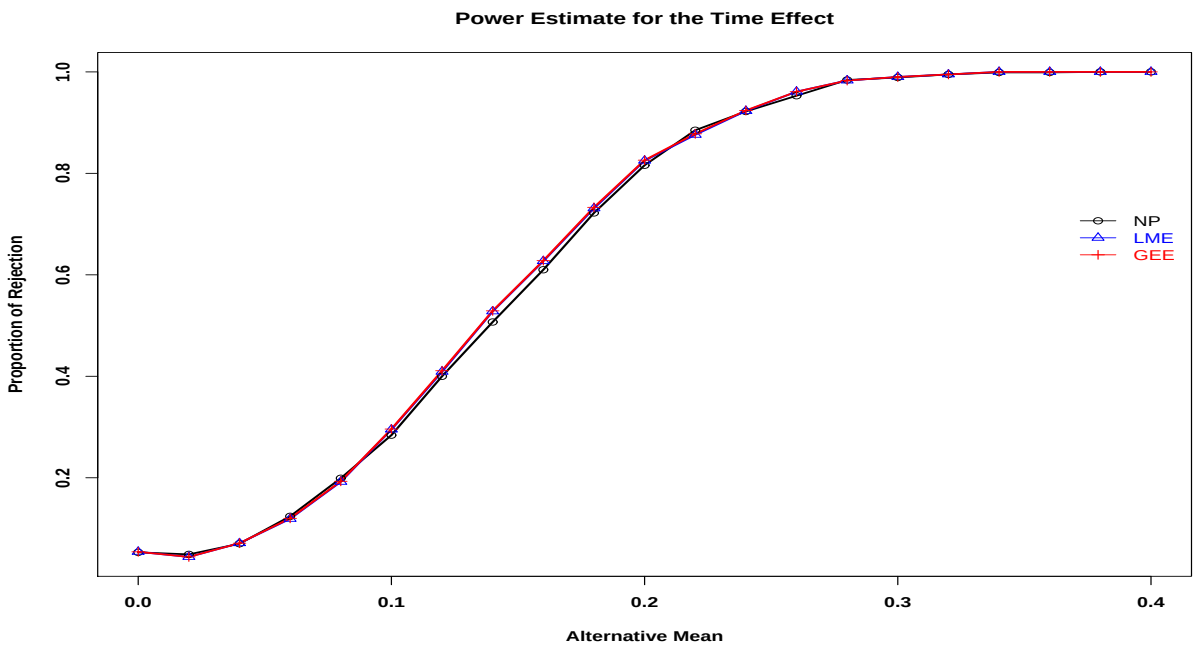


Figure 3.8: *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

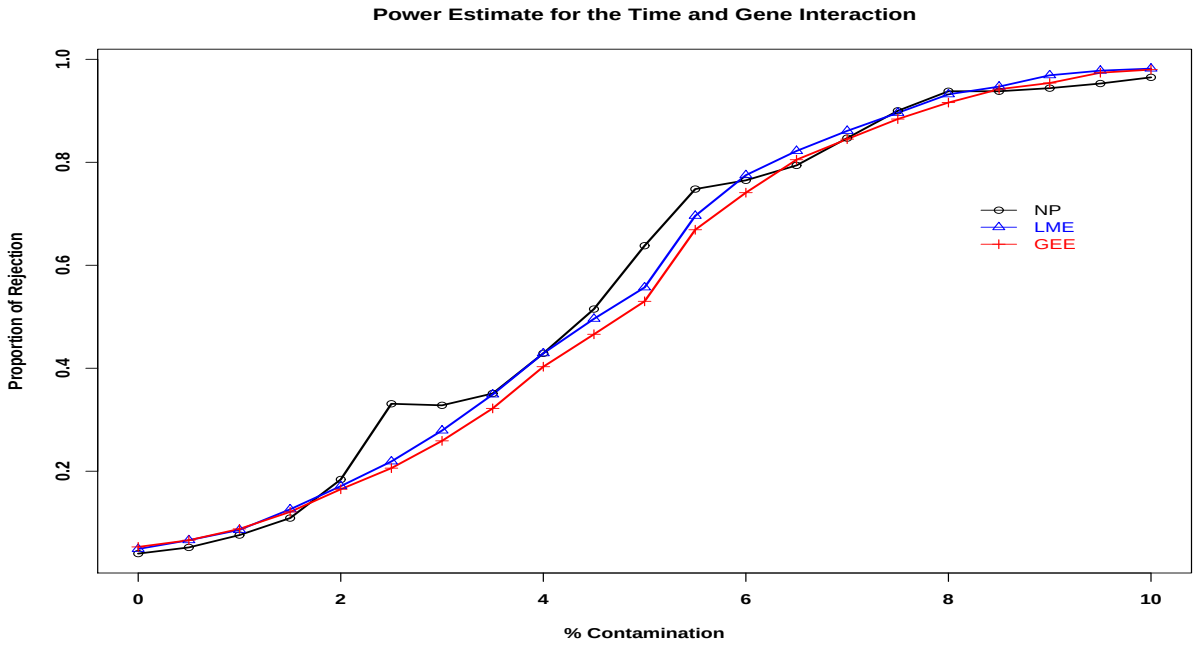


Figure 3.9: *The power curves of testing time effect with up to 1.5% contamination. The design was assumed to be unbalanced with unstructured correlation correlation. There were 2 time points in the experiment.*

microarray applications.

3.6 Real data analysis

Cytotoxic T lymphocyte (T cells) is of key importance in cell-mediated immune response. They destroy virally infected cells, tumor cells, and other disease cells. The effective immune response to a foreign antigen depends on rapid activation and proliferation of T cells. Interleukin-2 (IL-2) cytokine plays an important role in stimulating the growth, differentiation and survival of antigen-selected cytotoxic T cells via the activation of the expression of specific genes (Beadling et al. (1993)). A number of studies have been conducted to illustrate the gene expression profile with IL-2 stimulation and approximately 3000 IL-2-regulated genes were identified in human T cells (Beadling and Smith (2002); Kovanen et al. (2005); Mzali et al. (2005); Gatzka et al. (2006); Kovanen et al. (2008)).

Time course microarray study was recently carried out in Sandia National Laboratories to explore the expression profiles of IL-2 regulated genes during T cell proliferation and differentiation (Zhang et al. (2007)). The murine T cell line CTLL-2 was cultured in the presence or in the absence (control) of IL-2 stimulation. There were 3 independent cell cultures for either treatment group. For each culture, cells were harvested at time points 4 h and 8 h. Harvested cell samples were applied to microarray experiment with one array per sample. The Affymetrix Mouse Genome 430 2.0 Array were used. It comprises 45,000 probes representing approximately 30,000 mouse genes. We calculated the gene expression profile for each array by averaging the multiple probes of the gene. The gene profiles were log-transformed, and were then normalized with quantile-normalization method (Bolstad et al. (2003)). The normalized data were analyzed by gene set enrichment approach with the proposed statistics.

We used the C2 collection of gene sets from the Molecular Signature Database (MSigDB) of Broad Institute. C2 collection are curated gene sets that came from various sources such as

online pathway database, biomedical literature, and knowledge of domain experts (Newman and Weiner (2005)). The collection contains 1892 gene sets. The previous simulation studies have shown that a gene set consisting at least 50 genes would achieve sufficient statistical power and satisfactory type I error rate. Of the 1892 gene sets, 548 sets consist at least 50 genes. The distribution of the number of genes from the 548 gene sets was shown in Figure 3.10. The majority of gene sets (> 300) have between 50 and 100 genes. The largest gene set consists of 1601 genes. They are the genes enriched in mouse neural stem cell comparing to differentiated brain and marrow cells (Ramalho-Santos et al. (2002)).

In order to identify the gene sets that are regulated by IL-2, we used NPT to perform hypotheses testing for two effects, the interactions of treatment and time, and the main effect of IL-2 treatment. It is tempting to carry out test for the interaction of treatment and gene, for it would detect the gene sets that have a subset of genes differentially expressed between treatment groups. However, gene set enrichment analysis is only concerned with detecting expression alteration for the whole gene set, because a subset of genes would overlap with other gene sets and it does not convey all genetic information for an independent biological function. Therefore, we were not interested in such kind of gene sets. The P value of each test was converted to false discovery rate (FDR) with Storey's positive FDR method (Storey (2002)). The FDR was calculated by R package *fdrtool* (Strimmer (2008)).

With FDR cut off value at 5%, 285 out of 548 gene sets showed significant treatment*time interaction. In other words, they had differential expression between two treatment groups at 4 or 8 or both hours after IL-2 stimulation, but their expression patterns were distinct between the two time points. The biological analysis of the 285 gene sets need to be further explored. Of the remaining 263 gene sets, statistical tests for the treatment effect were performed, and 20 sets were identified to be significantly differentially expressed. Thus, all together we have 283 gene sets that are responsible to IL-2. The 20 selected gene sets for the treatment effect were reported in Table (3.8). Their FDR values were displayed as well. There were 1,760 distinct genes involved in the 20 gene sets.

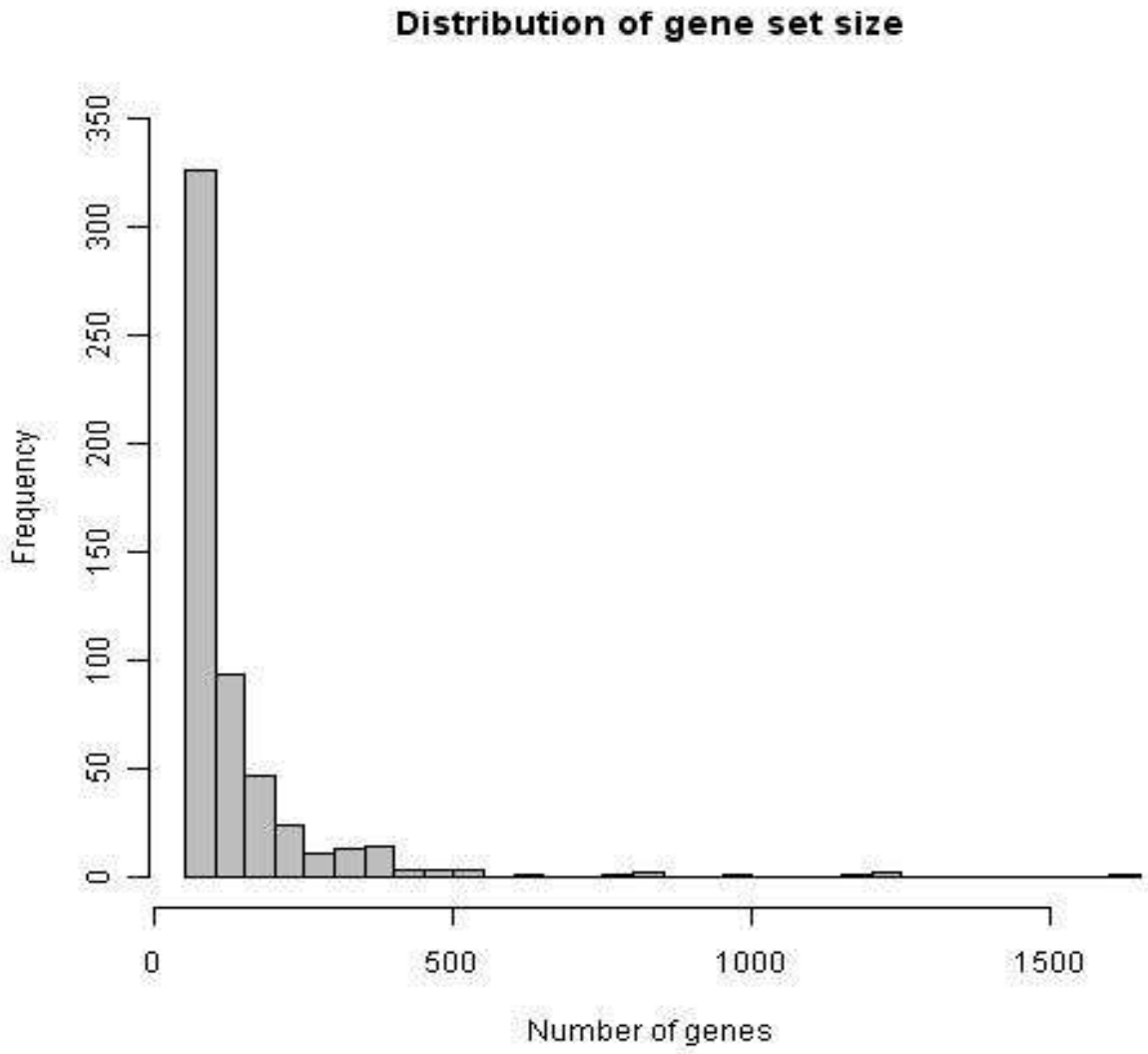


Figure 3.10: *The histogram showed the distribution of the size of the 548 gene sets used for data analysis.*

In order to illustrate how the gene expression in a selected gene set were uniformly altered by IL-2 over time period, we calculated *ratio score* by the formula: $ratio\ score = (log(G_{11}) - log(G_{12})) - (log(G_{21}) - log(G_{22}))$, where G_{ij} is the gene expression value at the ith treatment group, and at the jth time point. We expected that the ratio scores in the 20 selected gene sets to be close to zero. The ratio scores for each gene sets were plotted in Figures 3.11 and 3.12. They were distributed around the horizontal line at y=0, implying NPT selected the desired gene sets.

Gene Set	FDR
ROSS CBF	0.020
PEART HISTONE UP	0.047
ROME INSULIN 2F UP	0.038
HIVNEFPATHWAY	0.025
CELL ADHESION	0.041
HADDAD HSC CD7 UP	0.010
FLECHNER KIDNEY TRANSPLANT REJECTION PBL UP	0.009
SHEPARD POS REG OF CELL PROLIFERATION	0.029
HADDAD CD45CD7 PLUS VS MINUS UP	0.010
HSIAO LIVER SPECIFIC GENES	0.031
TAKEDA NUP8 HOXA9 3D UP	0.030
CROMER HYPOPHARYNGEAL MET VS NON DN	0.028
VANASSE BCL2 TARGETS	0.006
GAMMA UNIQUE FIBRO DN	0.018
TNFALPHA ADIP DN	0.026
GN CAMP GRANULOSA DN	0.041
AGED MOUSE NEOCORTEX UP	0.026
ADIP DIFF UP	0.006
HSA04370 VEGF SIGNALING PATHWAY	0.016
HSA04520 ADHERENS JUNCTION	0.008

Table 3.8: The IL-2 regulated gene sets.

T lymphocyte activation with IL-2 culminates many cellular processes, including blastogenesis, cell cycle progression, DNA replication and Mitosis (Beadling and Smith (2002)). Many of the selected gene sets are responsible such complicated biological functions. The gene set, VANASSE BCL2 TARGETS, consists of genes differentially expressed in murine CD19+ B cells overexpressing Bcl-2, a key gene regulating apoptosis. IL-2 is known to have antiapoptotic effects that proliferate T cells (Lenardo et al. (1999)). The other gene sets having similar functions of cell proliferation and aging are SHEPARD POS REG OF CELL PROLIFERATION, GAMMA UNIQUE FIBRO DN, and AGED MOUSE NEOCORTEX UP. Some selected gene sets, such as FLECHNER KIDNEY TRANSPLANT REJECTION PBL UP and HSIAO LIVER SPECIFIC GENES, are involved in the immune response of T cell. The gene sets, HADDAD HSC CD7 UP and HADDAD CD45CD7 PLUS VS MINUS

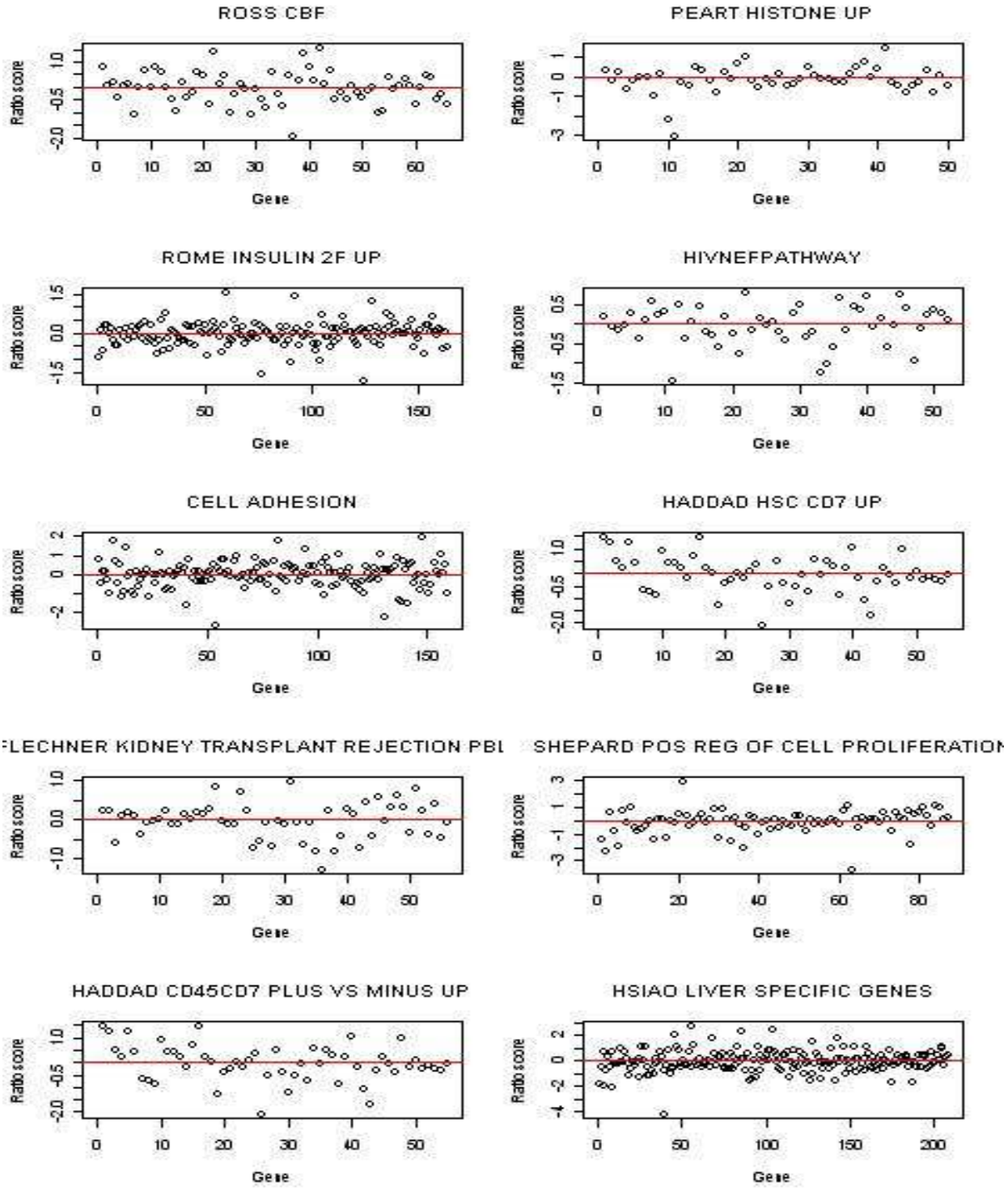


Figure 3.11: *The plots of log odds ratio for the first 10 selected gene sets.*

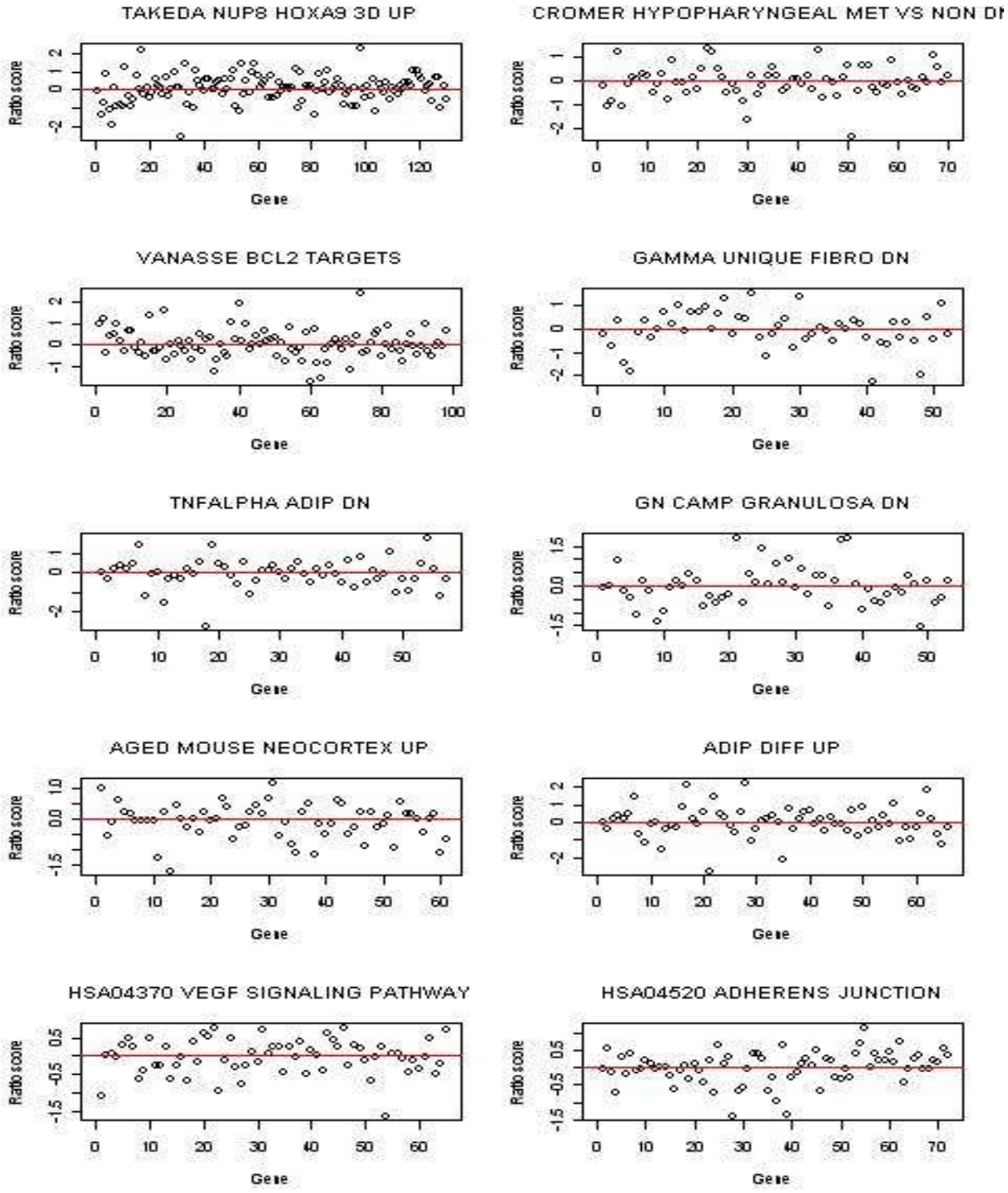


Figure 3.12: *The plots of log odds ratio for the second 10 selected gene sets.*

UP, are involved in T cell development. The gene sets, CELL ADHESION and HSA04520 ADHERENS JUNCTION, are responsible to the interaction of T cell with foreign cell, the core function of T cell mediated cytotoxicity. Insulin 2F related gene set, ROME INSULIN 2F UP, plays multiple roles in many gene regulating pathways including cell proliferation. The gene set HSA04370 VEGF SIGNALING PATHWAY often functions in tumor agiogenesis. The relationship of these gene sets with IL-2 stimulation is worth further investigation.

Most microarray data analyses in biological literature employ univariate analysis for each individual gene or probe. A list of genes is selected by FDR, and pathway analysis is then conducted for the candidate genes. On one hand, such analysis suffers low statistical power of detecting desired genes. On the other hand, the reported pathways are often misleading because they are based on one or very few selected genes, and those genes are most likely involved in many pathways. The proposed NPT methods via gene set enrichment analysis provides a promising alternative for biological functional analysis. The reported gene sets are highly relevant to the key functions of IL-2. It suggests high performance of the proposed methods in longitudinal microarray analysis.

Chapter 4

Rank-based Hypothesis Testing in Unbalanced Heteroscedastic Nested Design

In this chapter, we consider statistical testing for high dimensional data from a nested design when the number of lower-level factors is large. The proposed methods have potential applications in biological and meterological studies.

4.1 Model specification

Consider the nested design model:

$X_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$, i = 1, ..., a; j = 1, ..., b_i; k = 1, ..., c_{ij},

where α and β are fixed effects due to factor A and B, respectively. e_{ijk} is the error term with mean zero and variance $\sigma_{ij} < \infty$. And we assume $E({e_{ijk}}^4) = \delta_{ij} < \infty$. We consider the case that a and c_{ij} are fixed and $b_i \to \infty$.

Let

$X_{ijk} \sim F_{ij}$, i = 1, ..., a; j = 1, ..., b_i; k = 1, ..., c_{ij}

We have the decomposition

$$F_{ij} = M + A_i + B_{ij},$$

where

$$\sum_{i=1}^{a} A_i = \sum_{i=1}^{a} B_{ij} \sum_{j=1}^{b_i} B_{ij} = 0$$

We use the following notations in this chapter. X^* denotes a monotone transformation of X. $\widetilde{X}$ denotes the average of $\overline{X}_{ij\cdot}$'s.

4.2 Test statistics

To test the hypothesis that there is no main effect for A factor, i.e. $H_0(A)$: $A_i = 0$ for all i, we use a Wald-type test statistics

$$Q_{x^*}(A) = W' C_A' (C_A \hat{V} C_A')^{-1} C_A W, \tag{4.2.1}$$

where $W = (\tilde{X}^*_{1\cdot\cdot}, ..., \tilde{X}^*_{a\cdot\cdot})'$, $C_A = (\mathbf{1}_{a-1} j - I_{a-1})$, $\hat{V} = diag(\hat{\eta}_1, ..., \hat{\eta}_a)$, and

$$\hat{\eta}_i = \frac{1}{b_i^2} \sum_{j=1}^{b_i} \frac{S^2_{ij,X^*}}{c_{ij}},$$

with

$$S^2_{ij,X^*} = (c_{ij} - 1)^{-1} \sum_{k=1}^{c_{ij}} (X^*_{ijk} - \bar{X}^*_{ij\cdot})^2.$$

To test $H_0(B)$: $B_{ij} = 0$ for all i and j, we use a modified F test.

$$F_{X^*}(B) = \frac{MST_B}{MSE},$$

where

$$MST_B = \frac{1}{\sum_{i=1}^{a} b_i - a} \sum_{i=1}^{a} \sum_{j=1}^{b_i} (\tilde{X}^*_{ij\cdot} - \tilde{X}^*_{i\cdot\cdot})^2,$$

$$MSE = \frac{1}{a} \sum_{i=1}^{a} \frac{1}{b_i} \sum_{j=1}^{b_i} \frac{1}{c_{ij}(c_{ij}-1)} \sum_{k=1}^{c_{ij}} (X^*_{ijk} - \bar{X}^*_{ij\cdot})^2 = \frac{1}{a} \sum_{i=1}^{a} \frac{1}{b_i} \sum_{j=1}^{b_i} \frac{S^2_{ij,x}}{c_{ij}},$$

and n is the total number of samples.

4.3 Main results based on original observations

First, we consider the balanced case such that $b_i = b$ for all i. We have

$$MST_B = \frac{1}{ab - a} \sum_{i=1}^{a} \sum_{j=1}^{b} (\tilde{X}_{ij\cdot} - \tilde{X}_{i\cdot\cdot})^2,$$

and

$$MSE = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{1}{c_{ij}(c_{ij}-1)} \sum_{k=1}^{c_{ij}} (X_{ijk} - \bar{X}_{ij\cdot})^2 = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{S^2_{ij,x}}{c_{ij}}.$$

Theorem 4.3.1. *For testing $H_0(A)$: $A_i = 0$ for i=1,...,a, let $Q_X(A)$ be the statistic given in (4.2.1). If X_{ijk} has the finite fourth moment, then under $H_0(A)$,*

$$Q_X(A) \xrightarrow{d} \chi^2_{a-1}$$

holds for all $n_i \geq 2$, i=1, ..., I.

The proof is similar to that of theorem 3.2.1 in Wang (2004).

Theorem 4.3.2. *For testing $H_0(B)$, suppose X_{ijk} have finite variance σ_{ij}^2 and*

$$\limsup(ab)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{c_{ij}}E[X_{ijk}-E(X_{ijk})]^4<\infty.$$

Set

$$\tau_B=\frac{2}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_{ij}^4}{c_{ij}(c_{ij}-1)},$$

and

$$\sigma^2=\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_{ij}^2}{c_{ij}}$$

As $b\to\infty$,

$$\frac{\sqrt{ab}(F_x(B)-1)}{\sqrt{\tau_B}/\sigma^2}\xrightarrow{d}N(0,1)$$

holds when c_{ij} >3 stay fixed.

Proof: By Lemma A.1 And A.2, we only need to consider the asymptotic distribution of

$$T_B(e)=n(a,b)\sqrt{ab}(P_B(e)-MSE)$$

under H_0(B), where n(a,b) = $\min_{i,j}\{n_{ij}\}$, and

$$P_B(e)=\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\bar{e}_{ij\cdot}^2.$$

We have

$$
\begin{aligned}
T_B(e) &= n(a,b)\sqrt{ab}(P_B(e) - MSE) \\
&= \frac{n(a,b)}{\sqrt{ab}} \sum_{i=1}^{a} \sum_{j=1}^{b} \left[\bar{e}_{ij\cdot}^2 - \sum_{k=1}^{c_{ij}} \frac{(e_{ijk} - \bar{e}_{ij\cdot})^2}{c_{ij}(c_{ij}-1)} \right] \\
&= \frac{n(a,b)}{\sqrt{ab}} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k \neq k'}^{c_{ij}} \frac{e_{ijk} e_{ijk'}}{c_{ij}(c_{ij}-1)}.
\end{aligned}
$$

It is clear that $\mathrm{E}[\mathrm{T}_B(\mathrm{e})] = 0$, and as $\mathrm{b} \to \infty$,

$$
\begin{aligned}
Var[T_B(e)] &= \frac{n^2(a,b)}{ab} Var \left[\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k \neq k'}^{c_{ij}} \frac{e_{ijk} e_{ijk'}}{c_{ij}(c_{ij}-1)} \right] \\
&= \frac{2n^2(a,b)}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\sigma_{ij}^4}{c_{ij}(c_{ij}-1)} \\
&< \infty.
\end{aligned}
$$

We will use Lyapounov's theorem to obtain the asymptotic distribution of $\mathrm{T}_{1A}(\mathrm{e})$. Lyapounov's condition will be satisfied if

$$
L(b) = \sum_{j=1}^{b} E \left| \frac{n(a,b)}{\sqrt{ab}} \sum_{i=1}^{a} \sum_{k \neq k'}^{c_{ij}} \frac{e_{ijk} e_{ijk'}}{c_{ij}(c_{ij}-1)} \right|^4 \to 0.
$$

$$
\begin{aligned}
& L(b) \\
= & \sum_{j=1}^{b} E \left| \frac{n(a,b)}{\sqrt{ab}} \sum_{i=1}^{a} \sum_{k \neq k'}^{c_{ij}} \frac{e_{ijk} e_{ijk'}}{c_{ij}(c_{ij}-1)} \right|^4 \\
= & \sum_{j=1}^{b} \frac{n^4(a,b)}{(ab)^2} \sum_{i,i_1,i_2,i_3}^{a} \sum_{k \neq k'}^{c_{ij}} \sum_{k_1 \neq k_1'}^{c_{i_1 j}} \sum_{k_2 \neq k_2'}^{c_{i_2 j}} \sum_{k_3 \neq k_3'}^{c_{i_3 j}} \frac{E(e_{ijk} e_{ijk'} e_{i_1 jk_1} e_{i_1 jk_1'} e_{i_2 jk_2} e_{i_2 jk_2'} e_{i_3 jk_3} e_{i_3 jk_3'})}{c_{ij}(c_{ij}-1) c_{i_1 j}(c_{i_1 j}-1) c_{i_2 j}(c_{i_2 j}-1) c_{i_3 j}(c_{i_3 j}-1)} \\
= & O(\sum_{j=1}^{b} \frac{n^4(a,b)}{(ab)^2} \sum_{i,i_1}^{a} \sum_{k \neq k'}^{c_{ij}} \sum_{k_1 \neq k_1'}^{c_{i_1 j}} \frac{E(e_{ijk}^2) E(e_{ijk'}^2) E(e_{i_1 jk_1}^2) E(e_{i_1 jk_1'}^2)}{c_{ij}^2 (c_{ij}-1)^2 c_{i_1 j}^2 (c_{i_1 j}-1)^2}) \\
= & O(\frac{n^4(a,b)}{a^2 b^2} \sum_{j=1}^{b} (\sum_{i=1}^{a} \frac{\sigma_{ij}^4}{c_{ij}(c_{ij}-1)})^2) \\
= & O(b^{-1}).
\end{aligned}
$$

This completes the proof.

4.4 Main results based on ranks

In this section, we use the overall (mid-)ranks (R_{ijk}) of the original observations to test hypotheses. We denote $\mathrm{H(x)} = \mathrm{N}^{-1} \Sigma_{i,j} \mathrm{n}_{ij} \mathrm{F}_{ij}(\mathrm{x})$ and $\mathrm{Y}_{ijk} = \mathrm{H}(\mathrm{X}_{ijk})$, the average distribution function, and denote $\hat{H}$ the average of the left and right continuous version of the edf and

$$Z_{ijk} = \hat{H}(X_{ijk}).$$

Then, we have

$$R_{ijk} = N\hat{H}(X_{ijk}) + 0.5.$$

Theorem 4.4.1. *For testing $H_0(A)$: $A_i = 0$ for i=1,...,a, let $Q_R(A)$ be the statistic given*

in (4.2.1) with $X^*_{ijk} = R_{ijk}$. *If* X_{ijk} *has the finite fourth moment, then under* $H_0(A)$,

$$Q_R(A) \xrightarrow{d} \chi^2_{a-1}$$

holds for all $n_i \geq 2$, *i=1, ..., I.*

The proof is similar to that of theorem 3.4.1 in Wang (2004).

Theorem 4.4.2. *For testing* $H_0(B)$*: all* $B_j = 0$ *when a is fixed and b is large. Let* $F_R(B)$ *be the statistics* $F_{X^*}(B)$ *with* $X^*_{ijk} = R_{ijk}$. *As* $b \to \infty$,

$$\frac{\sqrt{ab}(F_R(B) - 1)}{\sqrt{\tau_B}/\sigma^2} \xrightarrow{d} N(0, 1),$$

where τ_B *and* σ^2 *are defined in Theorem 3.2 with evaluation at* Y_{ijk}.

Proof We denote MST_B(Y), MST_B(Z) and MST_B(R) the MST_B evaluated on Y, Z and R respectively. Note that $\mathrm{MST}_B(\mathrm{R})/\mathrm{N}^2 = \mathrm{MST}_B(\mathrm{Z})$. By lemma A.3 and A.4, it suffices to obtain the asymptotic distribution of

$$T(Z - E(Y)) = n(a, b)\sqrt{ab}(P_B(Z - E(Y)) - MSE(Z)).$$

We will show that T(Z - E(Y)) – T(Y – E(Y)) = $o_p(1)$ for

$$\frac{T(Y - E(Y))}{\sqrt{\tau_B}/\sigma^2} \xrightarrow{d} N(0, 1)$$

by theorem (4.3.2). The proof is similar to Lemma A.4 after the expand of T(Z - E(Y)) and T(Y – E(Y)) as in proof of theorem (4.3.2).

Appendix

Lemma A.1 Under the settings and assumptions of Theorem (4.3.2), we have n(a, b)(MSE - σ^2) $\to$ 0 in probability, where n(a,b) = $\min_{i,j}\{n_{ij}\}$.

Proof Note that

$$E(MSE) = E(\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{S_{ij,X}^2}{c_{ij}}) = \sigma^2,$$

and

$$\begin{aligned}
(ab)^2\mathrm{Var}(MSE) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{c_{ij}^2(c_{ij}-1)^2}\mathrm{Var}\left(\frac{c_{ij}-1}{c_{ij}}\sum_{k=1}^{c_{ij}}e_{ijk}^2 - \frac{1}{c_{ij}}\sum_{k\neq k'}^{c_{ij}}e_{ijk}e_{ijk'}\right) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{c_{ij}^4}\left[\sum_{k=1}^{c_{ij}}(E(e_{ijk}^4)-\sigma_{ij}^4) + \frac{2c_{ij}}{c_{ij}-1}\sigma_{ij}^4\right] \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{c_{ij}^4}\left[\sum_{k=1}^{c_{ij}}\delta_{ij}^4 - \frac{c_{ij}(c_{ij}-3)}{c_{ij}-1}\sigma_{ij}^4\right] \\
&\leq \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{c_{ij}^4}\sum_{k=1}^{c_{ij}}\delta_{ij}^4. \\
&\quad (c_{ij} > 3)
\end{aligned}$$

Thus

$$\mathrm{Var}[n(a,b)MSE] \leq \frac{1}{(ab)^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{n^2(a,b)}{n_{ij}^4}\sum_{k=1}^{n_{ij}}\delta_{ij}^4 \to 0$$

as b$\to\infty$. Therefore, n(a, b)(MSE - σ^2) $\to$ 0 in probability.

Lemma A.2 Define

$$P_B(e) = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\tilde{e}_{ij\cdot}^2.$$

Under the settings and assumptions of Theorem 3.2, and under H_0(B), we have

$$T_B^*(e) = n(a,b)\sqrt{ab}(MST_B - P_B(e)) \xrightarrow{P} 0$$

as b$\to\infty$, where n(a,b) = $\min_{i,j}\{c_{ij}\}$.

Proof Under H_0(B),

$$\begin{aligned} MST_B &= \frac{1}{a(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tilde{e}_{ij\cdot}-\tilde{e}_{i\cdot\cdot})^2 \\ &= \frac{1}{a(b-1)}\sum_{i=1}^{a}\left(\frac{b-1}{b}\sum_{j=1}^{b}\tilde{e}_{ij\cdot}-\frac{1}{b}\sum_{j\neq j'}^{b}\tilde{e}_{ij\cdot}\tilde{e}_{ij'\cdot}\right) \\ &= \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\tilde{e}_{ij\cdot}^2-\frac{1}{ab(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}\tilde{e}_{ij\cdot}\tilde{e}_{ij'\cdot}. \end{aligned}$$

Thus, we have

$$\begin{aligned} T_B^*(e) &= n(a,b)\sqrt{ab}\left(-\frac{1}{ab(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}\tilde{e}_{ij\cdot}\tilde{e}_{ij'\cdot}\right) \\ &= -\frac{n(a,b)}{\sqrt{ab}(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}\tilde{e}_{ij\cdot}\tilde{e}_{ij'\cdot}. \end{aligned}$$

It follows that

$$
\begin{aligned}
E[T_B^*(e)]^2 &= \frac{n^2(a,b)}{ab(b-1)^2} E\left[\sum_{i=1}^{a}\sum_{j\neq j'}^{b} \tilde{e}_{ij\cdot}\tilde{e}_{ij'\cdot}\right]^2 \\
&= \frac{2n^2(a,b)}{ab(b-1)^2} \sum_{i=1}^{a}\sum_{j\neq j'}^{b} E[\tilde{e}_{ij\cdot}^2\tilde{e}_{ij'\cdot}^2] \\
&= \frac{2n^2(a,b)}{ab(b-1)^2} \sum_{i=1}^{a}\sum_{j\neq j'}^{b} \frac{\sigma_{ij}^2}{c_{ij}}\frac{\sigma_{ij'}^2}{c_{ij'}} \\
&\leq \frac{2n^2(a,b)}{ab(b-1)} \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\sigma_{ij}^4}{c_{ij}} \\
&= O(b^{-1}).
\end{aligned}
$$

Therefore,

$$T_B^*(e) = n(a,b)\sqrt{ab}(MST_B - P_B(e)) \xrightarrow{P} 0$$

as b $\to \infty$.

Lemma A.3 Under the settings and assumptions of Theorem 4.2, we have n(a, b)(MSE/N^2 - σ^2) $\to$ 0 in probability as b $\to \infty$, where n(a,b) = $\min_{i,j}\{n_{ij}\}$ and a and c_{ij} remain fixed.

Proof For details, refer to the proof of Lemma 3.7.8 in Wang (2004).

Lemma A.4 Let P_B(Z – E(Y)) be defined as P_B(E) in lemma A.2 with e_{ijk} replaced by Z_{ijk} – E(Y_{ijk}). Then under H_0(B) and under the settings and assumptions of Theorem 4.2, as b $\to \infty$, we have

$$T_B^*(Z - E(Y)) = n(a,b)\sqrt{ab}(MST_B(Z) - P_B(Z - E(Y))) \xrightarrow{P} 0$$

as b$\to\infty$, where n(a,b) = $\min_{i,j}\{c_{ij}\} \geq 2$.

Proof By lemma A.2, T_B^*(Y - E(Y)) = $o_p(1)$ under $H_B(0)$. Therefore, it suffices to show that D_{ZY} = T_B^*(Z - E(Y)) - T_B^*(Y - E(Y)) = $o_p(1)$. Note that E(Z) = E(Y). Using the similar decompositions as in lemma A.2, under $H_B(0)$ we have

$$\begin{aligned} T_B^*(Y - E(Y)) &= n(a,b)\sqrt{ab}\left[-\frac{1}{ab(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}(\bar{y}_{ij\cdot} - \bar{p}_{ij\cdot})(\bar{y}_{ij'\cdot} - \bar{p}_{ij'\cdot})\right] \\ &= -\frac{n(a,b)}{\sqrt{ab}(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}(\bar{y}_{ij\cdot}\bar{y}_{ij'\cdot} - \bar{y}_{ij'\cdot}\bar{p}_{ij\cdot} - \bar{y}_{ij'\cdot}\bar{p}_{ij'\cdot} + \bar{p}_{ij\cdot}\bar{p}_{ij'\cdot}), \end{aligned}$$

and

$$\begin{aligned} T_B^*(Z - E(Y)) &= n(a,b)\sqrt{ab}\left[-\frac{1}{ab(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}(\bar{z}_{ij\cdot} - \bar{p}_{ij\cdot})(\bar{z}_{ij'\cdot} - \bar{p}_{ij'\cdot})\right] \\ &= -\frac{n(a,b)}{\sqrt{ab}(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}(\bar{z}_{ij\cdot}\bar{z}_{ij'\cdot} - \bar{z}_{ij'\cdot}\bar{p}_{ij\cdot} - \bar{z}_{ij'\cdot}\bar{p}_{ij'\cdot} + \bar{p}_{ij\cdot}\bar{p}_{ij'\cdot}). \end{aligned}$$

Then we have D_{ZY} = D_1 + D_2, where

$$D_1 = -\frac{n(a,b)}{\sqrt{ab}(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}(\bar{z}_{ij\cdot} - \bar{y}_{ij\cdot})(\bar{z}_{ij'\cdot} - \bar{y}_{ij'\cdot}),$$

and

$$\begin{aligned} D_2 &= -\frac{2n(a,b)}{\sqrt{ab}(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}(\bar{z}_{ij\cdot} - \bar{y}_{ij\cdot})(\bar{y}_{ij'\cdot} - \bar{p}_{ij'\cdot}) \\ &= -\frac{2n(a,b)}{\sqrt{ab}(b-1)}\sum_{i=1}^{a}\sum_{j\neq j'}^{b}\sum_{k,k'}\frac{(\bar{z}_{ijk} - \bar{y}_{ijk})(\bar{y}_{ij'k'} - \bar{p}_{ij'k'})}{c_{ij}c_{ij'}}. \end{aligned}$$

Because

$$\sup_x(\hat{H}(x) - H(x)) = O_p(N^{-1/2}),$$

we have

$$D_1 = O_p(\sqrt{ab}n(a,b)/N) = o_p(1).$$

Note $\mathrm{E}(\mathrm{D}_2) = 0$ for $\mathrm{E}(\mathrm{z}_{ijk}) = \mathrm{E}(\mathrm{y}_{ijk})$.

$$\begin{aligned}
E(D_2)^2 &= \frac{4n(a,b)^2}{ab(b-1)^2} \sum_{i=1}^{a} \sum_{j \neq j'}^{b} \sum_{k,k'} \frac{(\bar{z}_{ijk} - \bar{y}_{ijk})^2 (\bar{y}_{ij'k'} - \bar{p}_{ij'k'})^2}{c_{ij}^2 c_{ij'}^2} \\
&\leq \frac{4n(a,b)^2}{ab(b-1)^2} \sum_{i=1}^{a} \sum_{j \neq j'}^{b} \sum_{k,k'} \frac{1}{c_{ij}^2 c_{ij'}^2} \\
&= \frac{4n(a,b)^2}{ab(b-1)^2} \sum_{i=1}^{a} \sum_{j \neq j'}^{b} \frac{1}{c_{ij} c_{ij'}} \\
&\leq \frac{4}{b-1} \\
&= o_p(1).
\end{aligned}$$

This completes the proof.

Chapter 5

Summary and future studies

5.1 Summary of the current study

In recent years, high throughput technology have made available a great deal of biological data to the researchers. The technology has been increasingly applied to more complicated design such as time course study or nested design. The scientific community is in great need of statistical tools to analyze such high dimensional data. We attempted to provide a set of statistics for main effect tests for which traditional methodology is not successful. Robust statistics have been obtained for high dimensional data with heteroscadastic within subject correlation and unbalanced design. The asymptotic properties were provided as well. Comprehensive simulation studies have been conducted to test our methods in various conditions. In all conditions, our proposed methods provided satisfactory type I error rate. Power analysis were conducted to compare the approaches to linear mixed-effects model (LME) and generalized estimating equations (GEE). To allow the simulated data to closely represent real data, we used bootstrap to generate data based on applications from array CGH and expression microarray. The proposed methods were very sensitive with statistical testing power superior to that of LME and GEE in all tests. Our methods were applied to two recent longitudinal researches, Wilms' tumor aCGH study and IL-2 responsive microarray study. Comparing the literature reports, we provided results that were statistically more

justified and biologically more interpretable. It is promising to extend the methodology to a broader area of biological applications.

5.2 Future studies

Similar to the proposed methods, statistical methodology could be implemented and developed to a wide range of high dimensional data applications. We mainly consider three areas for future studies.

5.2.1 Spatially correlated image data

In Chapter 2, we focused on analysis longitudinal array CGH data. An important feature about array CGH is that probes are spatially correlated that adjacent probes tend to be deleted or amplified together. Similar spatial correlation is also observed in other image technology such as Magnetic Resonance Imaging (MRI) and Geographic Information Systems (GIS) imaging. Although most of current analyses of aCGH data assumed independence between probes (Weir et al. (2007); Sabatti and Lange (2008)), a few attempts have been made to consider the correlation. Fridlyand et al. (2004) used a hidden Markov model for the sequence of probes. However, it has to assume the correlation is exponentially distributed and the measures are log-normal distributed. Tibshirani and Wang (2008) proposed fused LASSO technique to constrain the copy number difference between neighboring probes. We implemented their method, but found it impossible to be applied to arrays with more than 100K probes due to the computation cost.

Wang et al. (2008) proposed a non-parametric clustering method for functional data whose time series satisfy an $\alpha - mixing$ condition. As the spatial sequence of probes bears similar correlation property to that of time series data, it is desirable to develop methodology for longitudinal aCGH study based on their techniques with proper consideration of correlations between time points.

5.2.2 Genetic interaction and gene networking

It is essential to elucidate gene-gene interaction for understanding how the basic biological activities of an organism are regulated by its genome. It has been a active research area to investigating genetic interaction and gene networking with high throughput technology (Brem et al. (2005); Zhong and Sternberg (2006)). In Chapter 3, we applied our test statistics to gene set enrichment analysis (GSEA) and detected activated gene groups. The selected genes within the same group are good candidates for investigating gene networking. Furthermore, our test statistics are potentially useful for test-based gene clustering and classification (Liao and Akritas (2007)).

5.2.3 High dimensional data integration

Nowadays, microarray or array CGH data are often collected from multiple centers or are acquired from various sources. It is important to control the batch effects between sources before data analysis. Furthermore, the multiple centers often use different platforms or versions of chips. Technically it is hard to integrate the data into a unified format (Irizarry et al. (2005)). In Chapter 4, we provided test statistics for nested design with high dimensional variables. They are intended to be used for high dimensional data integration. Simulation and real data anlaysis need be conducted to verify our methods.

Bibliography

Affymetrix (2006). *Affymetrix GeneChip chromosome copy number analysis tool (CNAT) - Version 4.0*. Affymetrix, Inc.

Beadling, C., K. W. Johnson, and K. A. Smith (1993). Isolation of interleukin 2-induced immediate-early genes. *Proc Natl Acad Sci U S A 90*(7), 2719–23.

Beadling, C. and K. A. Smith (2002). Dna array analysis of interleukin-2-regulated immediate/early genes. *Med Immunol 1*(1), 2.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*(2), 185–93.

Brem, R. B., J. D. Storey, J. Whittle, and L. Kruglyak (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature 436*(7051), 701–3.

Brennan, C., Y. Zhang, C. Leo, B. Feng, C. Cauwels, A. J. Aguirre, M. Kim, A. Protopopov, and L. Chin (2004). High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res 64*(14), 4744–8.

Brumback, B. and J. Rice (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc. 93*, 961–976.

Calvano, S. E., W. Xiao, Richards, R. M. D. R., Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, and S. K. Tschoeke (2005). A network-based analysis of systemic inflammation in humans. *Nature 437*, 1032–1037.

Carroll, R. and D. Ruppert (1988). *Transformation and Weighting in Regression*. Chapman and Hall.

Chen, M.-S., X. Liu, H. Wang, and M. Harris (2008). Hessian fly (mayetiola destructor) interactions with barley, rice, and wheat seedlings. *Submitted*.

Daruwala, R., A. Rudra, H. Ostrer, R. Lucito, M. Wigler, and B. Mishra (2004). A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci. USA 101*, 16292–16297.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of Longitudinal Data*. Oxford University Press, USA.

Dome, J. S., C. A. Bockhold, S. M. Li, S. D. Baker, D. M. Green, E. J. Perlman, D. A. Hill, and N. E. Breslow (2005). High telomerase rna expression level is an adverse prognostic factor for favorable-histology wilms' tumor. *J Clin Oncol 23*(36), 9138–45.

Du, J., J. Rozowsky, J. Korbel, Z. Zhang, T. Royce, M. Schultz, M. Snyder, and M. Gerstein (2006). A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and ChIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics 22*(24), 3016–3024.

Efron, B. and R. Tibshirani (2007). On testing the significance of sets of genes. *Annals of Applied Statistics 1*, 107–129.

Eggert, A., M. A. Grotzer, N. Ikegaki, H. Zhao, A. Cnaan, G. M. Brodeur, and A. E. Evans (2001). Expression of the neurotrophin receptor trkb is associated with unfavorable outcome in wilms' tumor. *J Clin Oncol 19*(3), 689–96.

Fan, J. and S. Lin (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc. 93*, 1007–1021.

Fan, J. and J.-T. Zhang (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal Of The Royal Statistical Society Series B 62*(2), 303–322.

Forozan, F., E. H. Mahlamaki, O. Monni, Y. Chen, R. Veldman, Y. Jiang, G. C. Gooden, S. P. Ethier, A. Kallioniemi, and O. Kallioniemi (2000). Comparative genomic hy-

bridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary dna microarray data. *Cancer Res. 60*, 4519–4525.

Fridlyand, J., A. Snijders, D. Pinkel, D. Albertson, and A. Jain (2004). Hidden markov models approach to the analysis of array CGH data. *J. Multivariate Anal. 90*, 132–153.

Gatzka, M., R. Piekorz, R. Moriggl, J. Rawlings, and J. N. Ihle (2006). A role for stat5a/b in protection of peripheral t-lymphocytes from postactivation apoptosis: insights from gene expression profiling. *Cytokine 34*(3-4), 143–54.

Gottardo, R., W. Li, W. Johnson, and X. Liu (2008). A flexible and powerful bayesian hierarchical model for ChIP-chip experiments. *Biometrics 64*(2), 468–78.

Gregorio, E. D., P. T. Spellman, G. M. Rubin, and B. Lemaitre (2001). Genome-wide analysis of the *drosophila* immune response by using oligonucleotide microarrays. *Proc. Natl. Acad. Sci. USA 98*, 2920–2929.

Greshock, J., T. L. Naylor, A. Margolin, S. Diskin, S. H. Cleaver, P. A. Futreal, P. J. deJong, S. Zhao, M. Liebman, and B. L. Weber (2004). 1-mb resolution array-based comparative genomic hybridization using a bac clone set optimized for cancer gene analysis. *Genome Res 14*(1), 179–87.

Grundy, P. E., N. E. Breslow, S. Li, E. Perlman, J. B. Beckwith, M. L. Ritchey, R. C. Shamberger, G. M. Haase, G. J. D'Angio, M. Donaldson, M. J. Coppes, M. Malogolowkin, P. Shearer, P. R. Thomas, R. Macklis, G. Tomlinson, V. Huff, and D. M. Green (2005). Loss of heterozygosity for chromosomes 1p and 16q is an adverse prognostic factor in favorable-histology wilms tumor: a report from the national wilms tumor study group. *J Clin Oncol 23*(29), 7312–21.

Guo, X., H. Qi, C. M. Verfaillie, and W. Pan (2003). Statistical significance analysis of longitudinal gene expression data. *Bioinformatics 19*(13), 1628–1635.

Hatchett, J., K. Starks, and J. Webster (1987). Insect and mite pests of wheat. *In Wheat and Wheat improvement, Agronomy Monograph 13*, 625–675.

Hing, S., Y. J. Lu, B. Summersgill, L. King-Underwood, J. Nicholson, P. Grundy, R. Grundy, M. Gessler, J. Shipley, and K. Pritchard-Jones (2001). Gain of 1q is associated with adverse outcome in favorable histology wilms' tumors. *Am J Pathol 158*(2), 393–8.

Hodgson, G., J. H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, N. Nowak, D. G. Albertson, D. Pinkel, C. Collins, D. Hanahan, and J. W. Gray (2001). Genome scanning with array cgh delineates regional alterations in mouse islet carcinomas. *Nat Genet 29*(4), 459–64.

Hoyle, D. C., M. Rattray, R. Jupp, and A. Brass (2002). Making sense of microarray data distributions. *Bioinformatics 18*(4), 576–84.

Hsu, L., S. G. Self, D. Grove, T. Randolph, K. Wang, J. J. Delrow, L. Loo, and P. Porter (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics 6*(2), 211–26.

Huang, S.-Y. and H. H.-S. Lu (2000). Bayesian wavelet shrinkage for nonparametric mixed-effects models. *Statistica Sinica 10*(4), 1021–1040.

Huang, S.-Y. and H. H.-S. Lu (2001). Extended gauss-markov theorem for nonparametric mixed-effects models. *Journal of Multivariate Analysis 76*(2), 249–266.

Hupe, P., N. Stransky, G. Thiery, F. Radvanyi, and E. Barillot (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics 20*, 3413–3422.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Res 31*(4), e15.

Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen,

J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu (2005). Multiple-laboratory comparison of microarray platforms. *Nat Methods 2*(5), 345–50.

Kalapurakal, J. A., J. S. Dome, E. J. Perlman, M. Malogolowkin, G. M. Haase, P. Grundy, and M. J. Coppes (2004). Management of wilms' tumour: current practice and future goals. *Lancet Oncol 5*(1), 37–46.

Kennedy, G. C., H. Matsuzaki, S. Dong, W. M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M. S. Phillips, M. T. Boyce-Jacino, S. P. Fodor, and K. W. Jones (2003). Large-scale genotyping of complex dna. *Nature Biotech. 21*, 1233–11237.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika 69*, 19–27.

Kerr, M. K., M. Martin, and G. A. Churchill (2000). Analysis of variance for gene expression microarray data. *J Comput Biol 7*(6), 819–37.

Konishi, T. (2004). Three-parameter lognormal distribution ubiquitously found in cdna microarray data and its application to parametric data treatment. *BMC Bioinformatics 5*, 5.

Kottgen, A., W. H. Kao, S. J. Hwang, E. Boerwinkle, Q. Yang, D. Levy, E. J. Benjamin, M. G. Larson, B. C. Astor, J. Coresh, and C. S. Fox (2008). Genome-wide association study for renal traits in the framingham heart and atherosclerosis risk in communities studies. *BMC Med Genet 9*, 49.

Kovanen, P. E., J. Bernard, A. Al-Shami, C. Liu, J. Bollenbacher-Reilley, L. Young, C. Pise-Masison, R. Spolski, and W. J. Leonard (2008). T-cell development and function are modulated by dual specificity phosphatase dusp5. *J Biol Chem 283*(25), 17362–9.

Kovanen, P. E., L. Young, A. Al-Shami, V. Rovella, C. A. Pise-Masison, M. F. Radonovich, J. Powell, J. Fu, J. N. Brady, P. J. Munson, and W. J. Leonard (2005). Global analysis of il-2 target genes: identification of chromosomal clusters of expressed

genes. *Int Immunol 17*(8), 1009–21.

Lai, L. A., T. G. Paulson, X. Li, C. A. Sanchez, C. Maley, R. D. Odze, B. J. Reid, and P. S. Rabinovitch (2007). Increasing genomic instability during premalignant neoplastic progression revealed through high resolution array-cgh. *Genes Chromosomes Cancer 46*(6), 532–42.

Lai, W., M. Johnson, R. Kucherlapati, and P. Park (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics 21*(19), 3763–3770.

Lenardo, M., K. M. Chan, F. Hornung, H. McFarland, R. Siegel, J. Wang, and L. Zheng (1999). Mature t lymphocyte apoptosis–immune regulation in a dynamic and unpredictable antigenic environment. *Annu Rev Immunol 17*, 221–53.

Li, W., C. Meyer, and X. Liu (2005). A hidden markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics 21*(Suppl 1:i), 274–82.

Liang, K.-Y. and S. L. Zeger (1986, April). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Liao, S.-M. and M. Akritas (2007). Test-based classification: A linkage between classification and statistical testing. *Statistics & Probability Letters 77*, 1269–1281.

Lu, Y. J., S. Hing, R. Williams, R. Pinkerton, J. Shipley, and K. Pritchard-Jones (2002). Chromosome 1q expression profiling and relapse in wilms' tumour. *Lancet 360*(9330), 385–6.

Maher, E. A. (2006, December). Marked genomic difference characterize primary and secondary gliobalstoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer Res 66*(23), 11502–11513.

Myers, C., M. Dunham, S. Kung, and O. Troyanskaya (2004). Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics 20*,

3533–3543.

Mzali, R., L. Seguin, C. Liot, A. Auger, P. Pacaud, G. Loirand, C. Thibault, J. Pierre, and J. Bertoglio (2005). Regulation of rho signaling pathways in interleukin-2-stimulated human t-lymphocytes. *Faseb J 19*(13), 1911–3.

Narita, S., N. Tsuchiya, L. Wang, S. Matsuura, C. Ohyama, S. Satoh, K. Sato, O. Ogawa, T. Habuchi, and T. Kato (2004). Association of lipoprotein lipase gene polymorphism with risk of prostate cancer in a japanese population. *Int J Cancer 112*(5), 872–6.

Natrajan, R., S. E. Little, J. S. Reis-Filho, L. Hing, B. Messahel, P. E. Grundy, J. S. Dome, T. Schneider, G. M. Vujanic, K. Pritchard-Jones, and C. Jones (2006). Amplification and overexpression of cacna1e correlates with relapse in favorable histology wilms' tumors. *Clin Cancer Res 12*(24), 7284–93.

Natrajan, R., S. E. Little, N. Sodha, J. S. Reis-Filho, A. Mackay, K. Fenwick, A. Ashworth, E. J. Perlman, J. S. Dome, P. E. Grundy, K. Pritchard-Jones, and C. Jones (2007). Analysis by array cgh of genomic changes associated with the progression or relapse of wilms' tumour. *J Pathol 211*(1), 52–9.

Newman, C. M. (1975, Septemper). An extension of khintchine's inequality. *Bulletin of the American Mathematical Society 81*(5), 913–915.

Newman, J. C. and A. M. Weiner (2005). L2l: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol 6*(9), R81.

Olshen, A., E. Venkatraman, R. Lucito, and M. Wigler (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics 5*, 557–572.

Olshen, A. B. and A. N. Jain (2002). Deriving quantitative conclusions from microarray expression data. *Bioinformatics 18*(7), 961–70.

Park, T., S.-G. Yi, S. Lee, S. Y. Lee, D.-H. Yoo, J.-I. Ahn, and Y.-S. Lee (2003). Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics 19*(6), 694–703.

Pinheiro, J. and D. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Pinkel, D. and D. G. Albertson (2005a). Array comparative genomic hybridization and its application in cancer. *Nature Genet. 37*, 511–517.

Pinkel, D. and D. G. Albertson (2005b). Comparative genomic hybridization. *Annu Rev Genomics Hum Genet 6*, 331–54.

Qi, H., D. J. Aguiar, S. M. Williams, A. L. Pean, W. Pan, and C. M. Verfaillie (2003). Identification of genes responsible for osteoblast differentiation from human mesodermal progenitor cells. *Proc. Natl. Acad. Sci. USA 100*, 2920–2929.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet 32 Suppl*, 496–501.

Ramalho-Santos, M., S. Yoon, Y. Matsuzaki, R. C. Mulligan, and D. A. Melton (2002). "stemness": transcriptional profiling of embryonic and adult stem cells. *Science 298*(5593), 597–600.

Ratcliffe, R., S. Cambron, K. Flanders, N. Bosque-Perez, S. Clement, and H. Ohm (2000). Biotype composition of hessian fly (diptera: Cecidomyiidae) populations from the southeastern, midwestern, and northwestern united states and virulence to resistance genes in wheat. *J. Econ. Entomol. 93*(4), 1319–1328.

Reiss, D., M. Facciotti, and N. Baliga (2007). Model-based deconvolution of genome-wide DNA binding. *Bioinformatics 24*(3), 396–403.

Rodwell1, G. E. J., R. Sonu, J. M. Zahn, J. Lund, J. Wilhelmy, L. Wang, W. Xiao, M. Mindrinos, E. Crane, E. Segal, B. D. Myers, J. D. Brooks, R. W. Davis, J. Higgins, A. B. Owen, and S. K. Kim (2004). A transcriptional profile of aging in the human kidney. *Plos Biol 2*, 2920–2929.

Sabatti, C. and K. Lange (2008). Bayesian gaussian mixture models for high-density genotyping arrays. *Journal of American Statistical Association 103*(481), 89–100.

Shedden, K. and S. Cooper (2002, July). Analysis of cell-cycle gene expression in saccharomyces cerevisiae using microarrays and multiple synchronization methods. *Nucleic Acids Res 30*(13), 2920–2929.

Sidorov, I. A., D. A. Hosack, D. Gee, J. Yang, M. C. Cam, R. A. Lempicki, and D. S. Dimitrov (2002). Oligonucleotide microarray data distribution and normalization. *Inf. Sci. Appl. 146*(1-4), 67–73.

Storey, J. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B 64*(3), 479–498.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA 100*(16), 9440–9445.

Storey, J. D., W. XIao, J. T. Leek, R. G. Tompkins, and R. W. Davis (2005). Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA 102*(36), 12837–12842.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics 9*, 303.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA 102*(43), 15545–15550.

Takahashi, M., X. J. Yang, T. T. Lavery, K. A. Furge, B. O. Williams, M. Tretiakova, A. Montag, N. J. Vogelzang, G. G. Re, A. J. Garvin, S. Soderhall, S. Kagawa, D. Hazel-Martin, A. Nordenskjold, and B. T. Teh (2002). Gene expression profiling of favorable histology wilms tumors and its correlation with clinical features. *Cancer Res 62*(22), 6598–605.

Tibshirani, R. and P. Wang (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics 9*(1), 18–29.

Tsai, G.-F. and A. Qu (2008). Testing the significance of cell-cycle patterns in time-course microarray data using nonparametric quadratic inference functions. *Comput. Stat. Data Anal. 52*(3), 1387–1398.

Tseng, G. C., M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong (2001). Issues in cdna microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res 29*(12), 2549–57.

Wahba, G. (1990). Spline models for observational data. *Regl. Conf. Ser. Appl. Math. 59*.

Wang, H. (2004). *Testing in multifactor heteroscedastic ANOVA and repeated measures designs with large number of levels*. Ph. D. thesis, Pennsulvania State University.

Wang, H. and M. Akritas (2004). Rank tests for anova with large number of factor levels. *Journal of Nonparametric Statistics 16*, 563–589.

Wang, H., J. Neill, and F. Miller (2008). Nonparametric clustering of functional data. *Statistics and its interface 1*, 47–62.

Wang, P., Y. Kim, J. Pollack, B. Narasimahan, and R. Tibshirani (2005). A method for calling gains and losses in array CGH data. *Biostatistics 6*, 45–58.

Wang, Y. (2002). Mixed effects smoothing spline analysis of variance. *J. R. Statist. Soc. B 60*(1), 159–174.

Weir, B. A., M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, K. Shah, M. Sato, R. K. Thomas, J. A. Barletta, I. B. Borecki, S. Broderick, A. C. Chang, D. Y. Chiang, L. R. Chirieac, J. Cho, Y. Fujii, A. F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B. E. Johnson, M. G. Kris, A. Lash, L. Lin, N. Lindeman, E. R. Mardis, J. D. McPherson, J. D. Minna, M. B. Morgan, M. Nadel, M. B. Orringer, J. R. Osborne, B. Ozenberger, A. H. Ramos, J. Robinson, J. A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M. R. Spitz, M. S. Tsao, D. Twomey, R. G. Verhaak, G. M. Weinstock, D. A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M. F. Zakowski, Q. Zhang, D. G. Beer, I. Wistuba,

M. A. Watson, L. A. Garraway, M. Ladanyi, W. D. Travis, W. Pao, M. A. Rubin, S. B. Gabriel, R. A. Gibbs, H. E. Varmus, R. K. Wilson, E. S. Lander, and M. Meyerson (2007). Characterizing the cancer genome in lung adenocarcinoma. *Nature 450*(7171), 893–8.

Williams, R. D., S. N. Hing, B. T. Greer, C. C. Whiteford, J. S. Wei, R. Natrajan, A. Kelsey, S. Rogers, C. Campbell, K. Pritchard-Jones, and J. Khan (2004). Prognostic classification of relapsing favorable histology wilms tumor using cdna microarray expression profiling and support vector machines. *Genes Chromosomes Cancer 41*(1), 65–79.

Yan, J. and J. Fine (2004). Estimating equations for association structures. *Stat Med 23*(6), 859–74; discussion 875–7,879–80.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res 30*(4), e15.

Yuan, E., C. M. Li, D. J. Yamashiro, J. Kandel, H. Thaker, V. V. Murty, and B. Tycko (2005). Genomic profiling maps loss of heterozygosity and defines the timing and stage dependence of epigenetic and genetic events in wilms' tumors. *Mol Cancer Res 3*(9), 493–502.

Zhang, Z., A. Martino, and J. L. Faulon (2007). Identification of expression patterns of il-2-responsive genes in the murine t cell line ctll-2. *J Interferon Cytokine Res 27*(12), 991–5.

Zhao, X., L. Cheng, J. G. Paez, and M. Meyerson (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res 64*, 3060–3071.

Zheng, M., L. Barrera, B. Ren, and Y. Wu (2007). ChIP-chip: data, model, and analysis. *Biometrics 63*, 787–796.

Zhong, W. and P. W. Sternberg (2006). Genome-wide prediction of c. elegans genetic interactions. *Science 311*(5766), 1481–4.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B 67*(2), 301–320.

Appendix A

R codes for data analysis

In this Appendix, we provide the R functions used for the simulation study and for the real data analysis. They are listed in the order of the chapters.

A.1 R functions for longitudinal aCGH study

For all the functions presented in this section, there are two input parameters. One is for the input data, named *Data*, *d*, or *sim.data*. The other is a vector of the number of replications for each probe, and it is named n or n_i in the following functions.

The input data should be a data matrix. Each row represents a time point, and each column represents a probe. Let X_{ijk} denotes the copy number of the ith probe, the jth time point, and the kth replicate. The input data matrix should be in the following format.

$$\begin{pmatrix} x_{111} & x_{112} & \cdot & \cdot & x_{211} & x_{212} & \cdot & \cdot & x_{a1n_a} \\ x_{121} & x_{122} & \cdot & \cdot & x_{221} & x_{222} & \cdot & \cdot & x_{a2n_a} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{1J1} & x_{1J2} & \cdot & \cdot & x_{2J1} & x_{2J2} & \cdot & \cdot & x_{aJn_a} \end{pmatrix}$$

where a is the number of probes, J is the number of time points, and n_a is the number of replications of the ath probe.

The vector of the numbers of the replications is in the format of $\{n_1, n_2, \cdots, n_a\}$, where

n_i is the number of replications for the ith probe.

For the functions used by NPT, the output value is the statistic calculated by NPT. For the functions of LME and GEE, a P value is reported.

A.1.1 R function for the test of the probe effect by NPT

```
# calculate the test statistic of the SNP effect for unbalanced data
calcStat.CN <- function(Data, n){

  b <- nrow(Data)
  a <- length(n)

  X<-Data



  VQ <- 0
  MSE <- 0
  for (i in 1:a) {
    if (i==1) start <- 1  else {
      start <- sum(n[1:(i-1)])+1
    }
    end <- sum(n[1:i])

    temp <- X[,start:end]
    temp.1 <- cbind(temp[,-1], temp[,1])
    Xd <- temp-temp.1 # paired difference X_{ijk}-X_{ijk+1}
```

```
  Xd.mult1 <- kronecker(Xd, rep(1,b))
  Xd.mult2 <- kronecker(rep(1,b), Xd)
  Xd.prod1 <- Xd.mult1 * Xd.mult2
  Xd.mult3 <- kronecker(rep(1,b^2), Xd.prod1)
  Xd.mult4 <- kronecker(cbind(Xd.prod1[,-c(1,2)],
               Xd.prod1[,c(1,2)]), rep(1, b^2))
  Xd.prod2 <- Xd.mult3 * Xd.mult4
  VQ <- VQ + sum(Xd.prod2)/(2*n[i]^2*(n[i]-1))

  # MSE
  Mean.1 <- apply(temp, 1, mean)  #mean of row
  Xm.d <- temp - Mean.1
  Xmd.mult1 <- kronecker(Xm.d,rep(1,b))
  Xmd.mult2 <- kronecker(rep(1,b),Xm.d)
  Xmd.prod <- Xmd.mult1 * Xmd.mult2
  MSE <- MSE + sum(Xmd.prod)/(n[i]*(n[i]-1))
}

TauA <- VQ /(a*b^2)
MSE <- MSE/(a*b)

ind <- rep(1:a, n)
Mean.B <- NULL  # average over replication n, b*a matrix
for (i in 1:b) {
  Mean.B <- rbind(Mean.B, tapply(X[i,], ind, mean))
}
```

```
  Mean.A <- apply(Mean.B, 2, mean)
  A <- Mean.A - mean(Mean.A)
  MST <- b / (a-1) * sum(A * A)

  Stat <- sqrt(a) * (MST - MSE) / sqrt(TauA)

  Stat

}
```

A.1.2 R function for the test of the time effect by NPT

```
# calculate the chi-sq statistic of time effect
# the unbalnaced design
calcStat.CN <- function(d, n.i){

  I <- length(n.i)
  J <- nrow(d)
  L <- rbind(t(rep(1, J-1)), -diag(J-1))

  #calculate means
  id.i <- rep(1:I, n.i)
```

```
  calcMeanij <- function(vd) tapply(vd, id.i, mean)
  mean.ij<-t(apply(d, 1, calcMeanij))



  mean.j <- apply(mean.ij,1,mean)

  #calculate eta
  data.d <-d - t(apply(mean.ij, 1, rep, times=n.i))  # x_ijk - mean.ij
  d1 <- kronecker(data.d, rep(1,J))
  d2 <- kronecker(rep(1,J), data.d)
  d.sq <- d1*d2
  d.sq.ij <- t(apply(d.sq, 1, calcMeanij)/(n.i-1))
  d.sq.j <- matrix(apply(d.sq.ij, 1, mean), J, J)  # matrix of length J^2

  eta <- t(L) %*% d.sq.j %*% L /I

  Stat <- t(mean.j) %*% L %*% solve(eta) %*% t(L) %*% mean.j
  Stat
}
```

A.1.3 R function for the test of the probe and time interaction by NPT

```
# calculate the test statistic of interaction effect for unbalanced data
calcStat.CN <- function(sim.data, n){

  b <- nrow(sim.data)
  a <- length(n)

  X<-sim.data

  # variance matrix V1 -- sum(\sigma^2_{i,jj1)) for any j, j1

  V1 <- 0
  V2 <- 0
  V3 <- 0
  MSE <- 0
  for (i in 1:a) {
    if (i==1) start <- 1    else {
      start <- sum(n[1:(i-1)])+1
    }
    end <- sum(n[1:i])

    temp <- X[,start:end]
    temp.1 <- cbind(temp[,-1], temp[,1])
    Xd <- temp-temp.1 # paired difference X_{ijk}-X_{ijk+1}

    Xd.mult1 <- kronecker(Xd, rep(1,b))
    Xd.mult2 <- kronecker(rep(1,b), Xd)
    Xd.prod <- Xd.mult1 * Xd.mult2
```

```
    V.mult1 <- kronecker(rep(1,b^2), Xd.prod)
    V.mult2 <- kronecker(cbind(Xd.prod[,-c(1,2)], Xd.prod[,c(1,2)]),
                   rep(1, b^2))
    V.prod <- V.mult1 * V.mult2

    #sigma^2 matrix
#    sigma.sq <- matrix(apply(Xd.prod, 1, mean)/2, J)

#    V1 <- V1 + sum(sigma.sq^2)/(n[i]*(n[i]-1))
    V1.id <- c(TRUE, rep(c(rep(FALSE, b^2), TRUE), b^2-1))
    V1 <- V1 + sum(V.prod[V1.id,])/(4*n[i]^2*(n[i]-1))

#    V2 <- V2 + sum(sigma.sq %*% sigma.sq)/(J^2*n[i]*(n[i]-1))
    V2 <- V2 + sum(V.prod)/(4*b^2*n[i]^2*(n[i]-1))

    V3.id <- c(rep(c(rep(TRUE, b), rep(FALSE, b^2-b)), b-1), rep(TRUE, b),
               rep(c(rep(FALSE, b), rep(c(rep(FALSE, b^2-b),
                       rep(TRUE, b)), b)), b-1))
    V3 <- V3 + sum(V.prod[V3.id,])/(2*b*n[i]^2*(n[i]-1))

    # MSE
    Mean.1 <- apply(temp, 1, mean)  #mean of row
    Xm.d <- temp - Mean.1
    Xmd.mult1 <- kronecker(Xm.d,rep(1,b))
    Xmd.mult2 <- kronecker(rep(1,b),Xm.d)
    Xmd.prod <- Xmd.mult1 * Xmd.mult2
```

```
    MSE <- MSE + sum(Xm.d*Xm.d)/(a*(b-1)*n[i]*(n[i]-1))
               - sum(Xmd.prod)/(a*b*(b-1)*n[i]*(n[i]-1))
  }

  TauA <- 2* (V1 + V2 - V3) /(a*(b-1)^2)
#  MSE <- MSE/(a*b)

  ind <- rep(1:a, n)
  Mean.AB <- NULL  # average over replication n, b*a matrix
  for (i in 1:b) {
    Mean.AB <- rbind(Mean.AB, tapply(X[i,], ind, mean))
  }

  Mean.A <- t(matrix(rep(apply(Mean.AB, 2, mean), b), ncol=b))
  Mean.B <- matrix(rep(apply(Mean.AB, 1, mean), a), nrow=b)
  Mean <- mean(Mean.A)

  AB <- Mean.AB - Mean.A - Mean.B + Mean
  MST <- 1 / ((a-1)*(b-1)) * sum(AB * AB)

  Stat <- sqrt(a) * (MST - MSE) / sqrt(TauA)

  Stat

}
```

A.1.4 Sample codes for LME and GEE calculations

We present two example functions for LME and GEE. They are used for the test of the probe and the time interaction. For the tests of the main effects, the codes need only to be slightly changed for the effect of interest.

```
## LME for the probe and time interactions
library(nlme)
calcStat.LME <- function(sim.data, n) {

  I <- length(n)
  J <- nrow(sim.data)

  Time <- as.vector(row(sim.data))
  SNP <- as.vector(t(matrix(rep(rep(1:I, n), J), ncol=J)))
  Sub <- as.vector(col(sim.data))
  CN <- as.vector(sim.data)
  X <- cbind(SNP, Time, Sub, CN)
  X <- data.matrix(X)

  gls.o=gls(CN~SNP+Time+SNP*Time, data=data.frame(X),
            corr=corSymm(form=~1|Sub))

  anova(gls.o, type="marginal")$"p-value"[4]
}
```

```
## GEE for the probe and time interaction.
library(geepack)
calcStat.GEE <- function(sim.data, n) {

  I <- length(n)
  J <- nrow(sim.data)

  Time <- as.vector(row(sim.data))
  SNP <- as.vector(t(matrix(rep(rep(1:I, n), J), ncol=J)))
  Sub <- as.vector(col(sim.data))
  CN <- as.vector(sim.data)
  X <- cbind(SNP, Time, Sub, CN)
  #X <- data.matrix(X)

  family <- "gaussian" #"poisson"
  gee.o=try(geese(CN~SNP+Time+SNP*Time, id=Sub,
              data=data.frame(X), family=family), T)

  if (!is(gee.o, "try-error")) geePvalue=c(summary(gee.o)$mean[4,4],
                1) else geePvalue=c(0,0)
  geePvalue     # pvalue for the trt effect
}
```

A.2 R functions for longitudinal microarray study with treatment groups

For all the functions presented in this section, there are two input parameters. One is for the input data, named *Data* or d. The other is a vector of the number of replications for each probe, and it is named n or n_{ik} in the following functions.

The input data should be a data matrix. Each row represents a time point, and each column represents a gene. Let X_{ijkl} denotes the copy number of the ith treatment group, the jth time point, and the kth gene, and the lth replicate. In the example of two treatment groups, the input data matrix should be in the following format.

$$\begin{pmatrix} x_{1111} & x_{1112} & \cdot & \cdot & x_{1121} & x_{1122} & \cdot & \cdot & x_{11Kn_{iK}} x_{2111} & x_{2112} & \cdot & \cdot & x_{2121} & x_{2122} & \cdot & \cdot & x_{21Kn_{iK}} \\ x_{1211} & x_{1212} & \cdot & \cdot & x_{1221} & x_{1222} & \cdot & \cdot & x_{12Kn_{iK}} x_{2211} & x_{2212} & \cdot & \cdot & x_{2221} & x_{2222} & \cdot & \cdot & x_{22Kn_{iK}} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{1J11} & x_{1J12} & \cdot & \cdot & x_{1J21} & x_{1J22} & \cdot & \cdot & x_{1JKn_{iK}} x_{2J11} & x_{2J12} & \cdot & \cdot & x_{2J21} & x_{2J22} & \cdot & \cdot & x_{2JKn_{iK}} \end{pmatrix}$$

where K is the number of genes, J is the number of time points, and n_{ik} is the number of replications of the kth probe in the ith treatment group.

The vector of the numbers of the replications is in the format of $\{n_{11}, n_{12}, \cdots, n_{1K}, n_{21}, n_{22}, \cdot\cdot\cdot, n_{2K}\}$, where n_{ik} is the number of replications for the kth gene in the ith treatment group.

For the functions used by NPT, the output value is the statistic calculated by NPT. For the functions of LME and GEE, a P value is reported.

In Chapter 3, we only considered the cases with 2 treatment groups for it is the most common. So in the sample codes, we assume two treatments group ($I = 2$).

A.2.1 R function for the test of the treatment effect by NPT

```
# calculate the test statistic of treatment effect.
```

```
calcStat.CN <- function(Data, n.ik){

  I <- 2
  J <- nrow(Data)
  K <- length(n.ik)/2

  #calculate means
  id.ik <- rep(1:(I*K),n.ik)
  mean.ijk<-NULL
  for (i in 1:J) {
    mean.ijk <- rbind(mean.ijk, tapply(Data[i,],id.ik,mean))
  }
  # mean.ijk <- apply(Data, 1, tapply, INDEX=id.ik, FUN=mean)

  mean.ik <- apply(mean.ijk,2,mean)

  id.i<-rep(1:I, rep(K,I))
  mean.i <- tapply(mean.ik, id.i, mean)

  #calculate eta
  Data.d <-Data - t(apply(mean.ijk, 1, rep, times=n.ik))
  d1 <- kronecker(Data.d, rep(1,J))
  d2 <- kronecker(rep(1,J), Data.d)
  d.sq <- d1*d2
  d.sq.ijk <- NULL
  for (i in 1:J^2) {
    d.sq.ijk <- rbind(d.sq.ijk, tapply(d.sq[i,], id.ik, sum))
```

```
  }
  d.sq.ijk <- t(t(d.sq.ijk)/(n.ik*(n.ik-1)))

  eta <- sum(d.sq.ijk)/J^2/K^2  #eta1+eta2

  Stat <- (mean.i[1]-mean.i[2])^2/eta
  Stat
}
```

A.2.2 R function for the test of the time effect by NPT

```
# calculate the chi-sq statistic of time effect.
calcStat.CN <- function(d, n.ik){
  I <- 2
  J <- nrow(d)
  K <- length(n.ik)/2

  L <- rbind(t(rep(1, J-1)), -diag(J-1))

  #calculate means
  ind.ijk <- rep(1:(2*K), n.ik)
  ind.jk <- rep(1:K, 2)
  mean.ijk <- NULL  # average over replication n, J*(2K) matrix
  for (j in 1:J) {
    mean.ijk <- rbind(mean.ijk, tapply(d[j,], ind.ijk, mean))
```

```
  }
  mean.jk <- NULL  # average over replication n, J*K matrix
  for (j in 1:J) {
    mean.jk <- rbind(mean.jk, tapply(mean.ijk[j,], ind.jk, mean))
  }
  mean.j <- apply(mean.jk,1,mean)  # J*1 vector

  #calculate eta=LVL'
  data.d <-d - t(apply(mean.ijk, 1, rep, times=n.ik))
  d1 <- kronecker(data.d, rep(1,J))
  d2 <- kronecker(rep(1,J), data.d)
  d.sq <- d1*d2
  d.sq.jk <- NULL
  for (j in 1:J^2) {
    d.sq.ijk <- tapply(d.sq[j,], ind.ijk, mean)/(n.ik-1)
    d.sq.jk <- rbind(d.sq.jk, tapply(d.sq.ijk, ind.jk, mean))
  }
  d.sq.j <- matrix(apply(d.sq.jk, 1, mean), J, J)

  eta <- t(L) %*% d.sq.j %*% L /I /K

  Stat <- t(mean.j) %*% L %*% solve(eta) %*% t(L) %*% mean.j
  Stat
}
```

A.2.3 R function for the test of the gene effect by NPT

```
# calculate the test statistic of gene effect for unbalanced data
calcStat.CN <- function(d, n.ik){
  I <- 2
  J <- nrow(d)
  K <- length(n.ik)/2
  n <- n.ik[1:K]

  X <- list(d[, 1:sum(n)], d[, (sum(n)+1):ncol(d)])

  V1 <- 0
  V2 <- 0
  MSE <- 0
  for (k in 1:K) {  # for each gene

    if (k==1) start <- 1     else {
      start <- sum(n[1:(k-1)])+1
    }
    end <- sum(n[1:k])

    sigma <- list(NULL, NULL)  # sigma estimation for 2 trts
    for (i in 1:I) {  # for each trt

      temp <- X[[i]][,start:end]
      temp.1 <- cbind(temp[,-1], temp[,1])
      Xd <- temp-temp.1 # paired difference X_{l}-X_{l+1}
```

```
    Xd.mult1 <- kronecker(Xd, rep(1,J))
    Xd.mult2 <- kronecker(rep(1,J), Xd)
    sigma[[i]] <- Xd.mult1 * Xd.mult2 / 2
    Xd.mult3 <- kronecker(rep(1,J^2), sigma[[i]])
    Xd.mult4 <- kronecker(cbind(sigma[[i]][,-c(1,2)],
                 sigma[[i]][,c(1,2)]), rep(1, J^2))
    Xd.prod2 <- Xd.mult3 * Xd.mult4
    V2 <- V2 + sum(Xd.prod2)/(n[k]^2*(n[k]-1))

    # MSE
    Mean.1 <- apply(temp, 1, mean)  #mean of row
    Xm.d <- temp - Mean.1
    Xmd.mult1 <- kronecker(Xm.d,rep(1,J))
    Xmd.mult2 <- kronecker(rep(1,J),Xm.d)
    Xmd.prod <- Xmd.mult1 * Xmd.mult2
    MSE <- MSE + sum(Xmd.prod)/(n[k]*(n[k]-1))
  }
  Xd.prod3 <- kronecker(sigma[[1]], rep(1, J^2)) * Xd.mult3
  V1 <- V1 + 2 * sum(Xd.prod3) / n[k]^3
}

TauA <- 2*(V1 + V2) /(I^2*J^2*K)
MSE <- MSE/(I*J*K)

# MST
ind <- rep(1:(2*K), n.ik)
```

```
  Mean.K <- NULL
  for (j in 1:J) {
    Mean.K <- rbind(Mean.K, tapply(d[j,], ind, mean))
  }

  Mean.J <- apply(Mean.K, 2, mean)  #1*2K matrix
  ind.i <- rep(1:K, 2)
  Mean.I <- tapply(Mean.J, ind.i, mean)  # 1*K matrix
  A <- Mean.I - mean(Mean.I)
  MST <- I*J / (K-1) * sum(A * A)

  Stat <- sqrt(K) * (MST - MSE) / sqrt(TauA)

  Stat

}
```

A.2.4 R function for the test of the treatment and time interaction by NPT

```
# calculate the chi-sq statistic of trt and time interaction.
calcStat.CN <- function(d, n.ik){
  I <- 2
  J <- nrow(d)
```

```
  K <- length(n.ik)/2

  L <- rbind(t(rep(1, I*J-1)), -diag(I*J-1))

  #calculate means
  ind.ijk <- rep(1:(2*K), n.ik)
#  ind.jk <- rep(1:K, 2)
  ind.ij <- c(rep(1,K), rep(2,K))
  mean.ijk <- NULL  # average over replication n, J*(2K) matrix
  for (j in 1:J) {
    mean.ijk <- rbind(mean.ijk, tapply(d[j,], ind.ijk, mean))
  }
  mean.ij <- NULL  # average over trts, J*I matrix
  for (j in 1:J) {
    mean.ij <- rbind(mean.ij, tapply(mean.ijk[j,], ind.ij, mean))
  }
  mean.ij.vec <- as.vector(mean.ij)
  mean.j <- apply(mean.ij,1,mean)  # J*1 vector

  #calculate eta=LVL'
  data.d <-d - t(apply(mean.ijk, 1, rep, times=n.ik))
  d1 <- kronecker(data.d, rep(1,J))
  d2 <- kronecker(rep(1,J), data.d)
  d.sq <- d1*d2        # (J^2)*(IK)
  d.sq.ij <- NULL      # (J^2)*I
  for (j in 1:J^2) {
    d.sq.ijk <- tapply(d.sq[j,], ind.ijk, mean)/(n.ik-1)
```

```
    d.sq.ij <- rbind(d.sq.ij, tapply(d.sq.ijk, ind.ij, mean))
  }
#  d.sq.j <- matrix(apply(d.sq.jk, 1, mean), J, J)  # J*J matrix
  d.sq.i <- diag(I*J)
  for (i in 1:I) {
    d.sq.i[((i-1)*J+1):(i*J), ((i-1)*J+1):(i*J)] <-
                    matrix(d.sq.ij[,i], nrow=J)
  }

  eta <- t(L) %*% d.sq.i %*% L /K  # (IJ)*(IJ)

  Stat <- t(mean.ij.vec) %*% L %*% solve(eta)
              %*% t(L) %*% mean.ij.vec
  Stat
}
```

A.2.5 R function for the test of the treatment and gene interaction by NPT

```
# calculate the test statistic of interaction effect of trt and gene
# for unbalanced data with unstructured correlation
calcStat.CN <- function(d, n.ik){
  I <- 2
  J <- nrow(d)
```

```
K <- length(n.ik)/2
n <- n.ik[1:K]

X <- list(d[, 1:sum(n)], d[, (sum(n)+1):ncol(d)])

V1 <- 0
V2 <- 0
MSE <- 0
for (k in 1:K) {  # for each gene

  if (k==1) start <- 1     else {
    start <- sum(n[1:(k-1)])+1
  }
  end <- sum(n[1:k])

  sigma <- list(NULL, NULL)  # sigma estimation for 2 trts
  for (i in 1:I) {  # for each trt

    temp <- X[[i]][,start:end]
    temp.1 <- cbind(temp[,-1], temp[,1])
    Xd <- temp-temp.1 # paired difference X_{l}-X_{l+1}

    Xd.mult1 <- kronecker(Xd, rep(1,J))
    Xd.mult2 <- kronecker(rep(1,J), Xd)
    sigma[[i]] <- Xd.mult1 * Xd.mult2 / 2
    Xd.mult3 <- kronecker(rep(1,J^2), sigma[[i]])
    Xd.mult4 <- kronecker(cbind(sigma[[i]][,-c(1,2)],
```

```
                    sigma[[i]][,c(1,2)]), rep(1, J^2))
    Xd.prod2 <- Xd.mult3 * Xd.mult4
    V2 <- V2 + sum(Xd.prod2)/(n[k]^2*(n[k]-1))

    # MSE
    Mean.1 <- apply(temp, 1, mean)  #mean of row
    Xm.d <- temp - Mean.1
    Xmd.mult1 <- kronecker(Xm.d,rep(1,J))
    Xmd.mult2 <- kronecker(rep(1,J),Xm.d)
    Xmd.prod <- Xmd.mult1 * Xmd.mult2
    MSE <- MSE + sum(Xmd.prod)/(n[k]*(n[k]-1))
  }
  Xd.prod3 <- kronecker(sigma[[1]], rep(1, J^2)) * Xd.mult3
  V1 <- V1 + 2 * sum(Xd.prod3) / n[k]^3
}

TauA <- 2*(V1/(I-1)^2 + V2) /(I^2*J^2*K)
MSE <- MSE/(I*J*K)

# MST
ind.ijk <- rep(1:(2*K), n.ik)
ind.i <- c(rep(1,K), rep(2,K))
ind.k <- rep(1:K, 2)
Mean.ijk <- NULL
for (j in 1:J) {
  Mean.ijk <- rbind(Mean.ijk, tapply(d[j,], ind.ijk, mean))
}
```

```
  Mean.ik <- apply(Mean.ijk, 2, mean)
  Mean.i <- tapply(Mean.ik, ind.i, mean)
  Mean.k <- tapply(Mean.ik, ind.k, mean)
  Mean <- mean(Mean.k)

  A <- Mean.ik - rep(Mean.i, rep(K,I)) - rep(Mean.k, I) + Mean
  MST <- J /((I-1) * (K-1)) * sum(A * A)

  Stat <- sqrt(K) * (MST - MSE) / sqrt(TauA)

  Stat

}
```

A.2.6 R function for the test of the gene and time interaction by NPT

```
# calculate the test statistic of interaction effect of time
# and gene for unbalanced data with unstructured correlation
calcStat.CN <- function(d, n.ik){

  I <- 2
  J <- nrow(d)
```

```
K <- length(n.ik)/2
n <- n.ik[1:K]

X <- list(d[, 1:sum(n)], d[, (sum(n)+1):ncol(d)])



V1 <- 0
V2 <- 0
V3 <- 0
MSE <- 0
for (k in 1:K) {  # for each gene

  if (k==1) start <- 1     else {
    start <- sum(n[1:(k-1)])+1
  }
  end <- sum(n[1:k])

  for (i in 1:I) {  # for each trt

    temp <- X[[i]][,start:end]
    temp.1 <- cbind(temp[,-1], temp[,1])
    Xd <- temp-temp.1 # paired difference X_{l}-X_{l+1}

    Xd.mult1 <- kronecker(Xd, rep(1,J))
    Xd.mult2 <- kronecker(rep(1,J), Xd)
    Xd.prod <- Xd.mult1 * Xd.mult2
```

```
    V.mult1 <- kronecker(rep(1,J^2), Xd.prod)
    V.mult2 <- kronecker(cbind(Xd.prod[,-c(1,2)],
                  Xd.prod[,c(1,2)]), rep(1, J^2))
    V.prod <- V.mult1 * V.mult2

    V1.id <- c(TRUE, rep(c(rep(FALSE, J^2), TRUE), J^2-1))
    V1 <- V1 + sum(V.prod[V1.id,])/(4*n[k]^2*(n[k]-1))

    V2 <- V2 + sum(V.prod)/(4*J^2*n[k]^2*(n[k]-1))

    V3.id <- c(rep(c(rep(TRUE, J), rep(FALSE, J^2-J)), J-1),
                rep(TRUE, J), rep(c(rep(FALSE, J), rep(c(rep(FALSE,
                J^2-J), rep(TRUE, J)), J)), J-1))
    V3 <- V3 + sum(V.prod[V3.id,])/(2*J*n[k]^2*(n[k]-1))

    # MSE
    Mean.1 <- apply(temp, 1, mean)  #mean of row
    Xm.d <- temp - Mean.1
    Xmd.prod1 <- Xm.d*Xm.d
    Xmd.mult1 <- kronecker(Xm.d,rep(1,J))
    Xmd.mult2 <- kronecker(rep(1,J),Xm.d)
    Xmd.prod2 <- Xmd.mult1 * Xmd.mult2
    MSE <- MSE + sum(Xmd.prod1)/(n[k]*(n[k]-1))
             - sum(Xmd.prod2)/(J*n[k]*(n[k]-1))

  }
}
```

```
  TauA <- 2* (V1 + V2 - V3) /(I*K*(J-1)^2)
  MSE <- MSE/(I*K*(J-1))

  # MST
  ind.ijk <- rep(1:(2*K), n.ik)
  ind.jk <- rep(1:K, 2)
  Mean.jk <- NULL
  for (j in 1:J) {
    Mean.ijk <- tapply(d[j,], ind.ijk, mean)
    Mean.jk <- rbind(Mean.jk, tapply(Mean.ijk, ind.jk, mean))
  }

  Mean.j <- apply(Mean.jk, 1, mean)
  Mean.k <- apply(Mean.jk, 2, mean)
  Mean <- mean(Mean.k)
  A <- Mean.jk - Mean.j - kronecker(t(Mean.k), rep(1,J)) + Mean
  MST <- I /((J-1) * (K-1)) * sum(A * A)

  Stat <- sqrt(K) * (MST - MSE) / sqrt(TauA)

  Stat

}
```

A.2.7 Sample codes for LME and GEE calculations

We present two example functions for LME and GEE. They are used for the test of the gene and the time interaction. For the tests of the main effects and other interactions, the codes need only to be slightly changed for the effect of interest.

```
## LME for the gene and time interactions
library(nlme)
calcStat.LME <- function(Data, n) {

  I <- 2
  J <- nrow(Data)
  K <- length(n)/2

  Trt <- as.vector(col(Data))
  Trt[Trt<=sum(n)/2] <- 1
  Trt[Trt>sum(n)/2] <- 2
  Time <- as.vector(row(Data))
  Gene <- as.vector(t(matrix(rep(c(rep(1:K, n[1:K]),
         rep(1:K, n[(K+1):(2*K)])), J), ncol=J)))
  Sub <- as.vector(col(Data))
  Exp <- as.vector(Data)
  X <- cbind(Trt, Gene, Time, Sub, Exp)
  X <- data.matrix(X)
```

```
  gls.o=gls(Exp~Trt+Gene+Time+Time*Gene,
    data=data.frame(X),  corr=corSymm(form=~1|Sub))

  anova(gls.o, type="marginal")$"p-value"[5]
}



## GEE for the gene and time interaction.
library(geepack)
calcStat.GEE <- function(Data, n) {

  I <- 2
  J <- nrow(Data)
  K <- length(n)/2

  Trt <- as.vector(col(Data))
  Trt[Trt<=sum(n)/2] <- 1
  Trt[Trt>sum(n)/2] <- 2
  Time <- as.vector(row(Data))
  Gene <- as.vector(t(matrix(rep(c(rep(1:K, n[1:K]),
          rep(1:K, n[(K+1):(2*K)])), J), ncol=J)))
  Sub <- as.vector(col(Data))
  Exp <- as.vector(Data)
  X <- cbind(Trt, Gene, Time, Sub, Exp)
```

```
  family <- "gaussian" #"poisson"
  gee.o=try(geese(Exp~Trt+Gene+Time+Time*Gene, id=Sub,
                data=data.frame(X), family=family), T)

  if (!is(gee.o, "try-error")) geePvalue=c(summary(
          gee.o)$mean[5,4],1) else geePvalue=c(0,0)
  geePvalue     # pvalue for the trt effect
}
```